Diversification of Sundaland aquatic biotas : build-up of freshwater fishes' diversity and distribution in a biodiversity hotspot

Arni Sholihah

▶ **To cite this version:**

Arni Sholihah. Diversification of Sundaland aquatic biotas : build-up of freshwater fishes' diversity and distribution in a biodiversity hotspot. Agricultural sciences. Université Montpellier, 2020. English. NNT : 2020MONTG024 . tel-03155262

HAL Id: tel-03155262
https://theses.hal.science/tel-03155262

Submitted on 1 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Sciences de l'Évolution et de la Biodiversité

École doctorale GAIA (N°584)

L'Institut des Sciences de l'Évolution de Montpellier (ISEM)

Diversification of Sundaland aquatic biotas: build-up of freshwater fishes' diversity and distribution in a biodiversity hotspot

Présentée par Arni SHOLIHAH
Le 14 Décembre 2020

Sous la direction de Jean-François AGNÈSE et Nicolas HUBERT

Devant le jury composé de

Nicolas PUILLANDRE, HDR, MC MNHN	Rapporteur
Mark DE BRUYN, Lecturer University of Sydney	Rapporteur
Philippe KEITH, HDR, Professeur MNHN	Examinateur
Emmanuel DOUZERY, HDR, Professeur Université de Montpellier	Examinateur/Président
Jean-François AGNÈSE, HDR, DR IRD	Directeur de thèse
Nicolas HUBERT, HDR, DR IRD	Co-directeur de thèse
Fabien L. CONDAMINE, CR CNRS	Invité
Lukas RÜBER, Curator Naturhistorisches Museum Bern	Invité

UNIVERSITÉ DE MONTPELLIER

Rincik rincang rincik rincang

Kai bobo di walungan

Sidik pisan sidik pisan

Si bogo teh, leutik 'ningan

Si Kabayan kakoncara

Néang paray meunang kancra

Ayeuna tinggal waas na

Walungan tiwas dipergasa

Baheula guyub katelahna

Nguseup beunteur 'clom giriwil

Ayeuna béak kasebutna

Lung na useup moal hasil

Beuleum tawés pais nilem

Ulah tinggaleun sambelna

Beureum panon mangkat tilem

Cenah ninggalkeun Nyi Sunda

Ngurek belut manggih lélé

Kadé tong keuna patilna

Loba gelut, mikir lénglé

Jalma loba ngaruksakna

Nyangkéré kéhkél na batu

Ngagogo sepat di sawah

Mangkadé isuk can tangtu

Ngagugu adat nu salah

Marak walungan Pasundan

Lauk Afrika meunangna

Harak ngagarong lingkungan

Burung sangsara ahirna…

Table of Contents

List of Figures

List of Tables

Abstract

Sundaland is one of the most threatened biodiversity hotspots, experiencing a fast increase of threat levels during last decades. Covering Malayan Peninsula, Sumatra, Java and Borneo, this hotspot has one of the highest species richness and endemism for vertebrates in SEA, including freshwater fishes. This level of biodiversity has long attracted the attention of evolutionary biologists, particularly by considering effects of Sundaland complex geological history. This study addressed it by exploring time frame of vicariance and dispersal during diversity build-up of freshwater fish species in Sundaland. To support this, we first aimed to assess the match between distribution of molecular lineages from multiple taxa with palaeoriver boundaries using metadata analysis of existing molecular dataset with representative biological and spatial coverage in Southeast Asia (especially in Sundaland). Second, we focussed on estimating clades' age and geographic distribution of *Rasbora* lineages in relation to the Pleistocene Palaeoriver Hypothesis by utilising newly generated empirical data for Rasborinae, a widespread and extremely diversified group of primary freshwater fishes in Sundaland. On both steps, we questioned: 1) if palaeorivers served as corridors of dispersal between islands during Pleistocene sea levels low stands; 2) if palaeoriver watersheds initiated allopatric divergence across their boundaries; and 3) if Pleistocene climatic fluctuation increased rates of species diversification. Overall, this study detected high level of cryptic diversity. Ancestral area reconstructions revealed that Sundaland freshwater fish lineages originated from Mainland Asia, and further colonised the region since Oligocene. This result validated the pre-Pleistocene settlement hypothesis. These lineages entered Sundaland mainly through North Sunda palaeoriver in contemporary Borneo and dispersed to other parts of Sundaland via long distance dispersal, often followed by *in situ* diversification. These results suggest Bornean part of North Sunda palaeoriver is the most likely centre of origin for Sundaland freshwater fishes. Contrary to the initial hypothesis, we found that although lowered sea level during glacial periods reconnected watersheds within palaeorivers, it did not necessarily open up inter-island dispersal channels for freshwater fishes. Corridors of savanna and seasonal forest ecosystems in the interior of Sundaland served as barrier to dispersal. Also, permeability of the physical boundaries of palaeoriver's watersheds as well as geomorphological and habitat variabilities within palaeoriver created respectively gene flow between palaeorivers and allopatric speciation within palaeoriver. Moreover, although significant proportion of Sundaland freshwater fish lineages originated during Pleistocene, we found that Pleistocene dynamics did not affect diversification rate as sea level-dependent diversification models poorly account for species proliferation patterns for all clades excepting *Channa*. Besides, none of the taxa examined has declining diversification rates as suggested by diversity-dependent diversification (DDD) model. It is suggested then that global Pleistocene eustatic fluctuation and regional paleoriver dynamics are not sole drivers for Sundaland freshwater fish diversification, but only a part of abiotic aspects affecting it. Pleistocene Climatic Fluctuations likely interacted with other factors such as: landscape geomorphology, local ecosystem/habitat variability and life history traits of organisms. [**Eustatic fluctuations**; **insular systems**; **diversification**; **vicariance**; **dispersal**; **palaeoenvironments**; **freshwater fishes; mitogenomes; DNA-based species delimitation**]

Résumé

En raison de l'histoire géologique et des dynamiques écologiques locales, la biodiversité n'est pas uniformément répartie sur Terre, à la fois dans le temps et dans l'espace. Sur le plan chronologique, certains groupes peuvent être majoritairement abondants pendant une certaine période et très rares ou même disparus à d'autres périodes. Pendant ce temps, la cartographie spatiale de la biodiversité au cours d'une même période de temps montre également un modèle de distribution très hétérogène. Dans ce contexte, la biogéographie, en particulier la phylogéographie, étudie les principes et les processus régissant le modèle de distribution géographique des lignées à travers le temps, comme le montre les inférences historiques basées sur les phylogénies moléculaires.

En plus de ces facteurs, les pressions humaines sur l'environnement ont également créé des changements importants de la biodiversité. Elle a actuellement tendance à s'accumuler dans les «points chauds de la biodiversité», des zones d'endémisme exceptionnel confrontées à des pressions anthropiques massives. Parmi les 36 points chauds de biodiversité recensés dans le monde, quatre se trouvent en Asie du Sud-Est (SEA), à savoir: Indo-Birmanie, Philippines, Sundaland et Wallacea. Parmi eux, le Sundaland est actuellement l'un des plus menacés au monde et celui qui a connu l'augmentation la plus rapide des niveaux de menace au cours des dernières décennies. Couvrant la péninsule malaise, Sumatra, Java et Bornéo, cette région se situe géologiquement sur le plateau de la Sonde. Sundaland connaît un climat tropical chaud avec une certaine hétérogénéité spatiale qui peut être attribuée à: différents substrat géologiques; drainage de l'eau; et la topographie qui comprend un large éventail de paysages allant des plaines d'inondation au niveau de la mer jusqu'aux chaînes de montagnes. Naturellement, les habitats du Sundaland couvrent de nombreux types d'écosystèmes, tels que: forêt tropicale (de montagne, de plaine); forêt marécageuse de tourbe; divers habitats d'eau douce comme les torrents de montagne, les ruisseaux, les rivières à faible débit, les lacs, les marécages, les plaines inondables des basses terres; les habitatx d'eau saumâtre (par exemple, forêts de mangroves); et des mers peu profondes riches en biodiversité marine.

D'un point de vue biogéographique, le schéma général de la biodiversité du Sundaland ressemble davantage à l'Asie du Sud-Est continentale par rapport à ses

voisins Wallacea et Papua (plateau du Sahul). Cette région a une grande richesse en espèces et une forte endémicité des plantes supérieures et des vertébrés. Par exemple, près de 900 des 1200 espèces de poissons d'eau douce d'Indonésie sont présente dans le Sundaland, et environ 400 d'entre elles sont endémiques. Couplé aux fortes pressions anthropiques liés au développement socio-économique local et régional, cette situation biologique place le Sundaland comme l'un des points chauds de biodiversité les plus menacés au monde. En outre, les efforts de conservation dans la zone sont également limités par la présence d'une importante diversité cryptique, y compris chez les biotas d'eau douce.

Ce niveau de biodiversité régionale a longtemps un centre d'attention de la biologie évolutive, notamment en considérant les effets de l'histoire géologique, dynamique et complexe, du Sundaland. En formation depuis le Cénozoïque en tant que pointe au sud de l'Eurasie, le Sundaland a subi des changements graduels à la suite de mouvements tectoniques, combinés en cours de route par l'émergence de nouvelles îles dans la région. Les mers régionales peu profondes autour du Sundaland ont commencé à s'ouvrir à la fin du Miocène (10 Ma) et la configuration contemporaine des masses terrestres du Sundaland est généralement stable depuis le début du Pliocène (5 Ma). Compte tenu de cette dynamique, il a été suggéré que le paléoclimat du Sundaland avait également changé avec le temps, impactant l'étendue des masses terrestres et des mers peu profondes. Le paléoclimat du Sundaland au cours du Mio-Pliocène (20-5 Ma) a été estimé comme plus chaud et plus humide, tandis qu'au Pléistocène, le paléoclimat pouvait différer en fonction de cycles glaciaires.

Le Sundaland étant entouré de mers peu profondes, dépassant rarement 100m de profondeur, l'étendue de ses masses continentales a put largement fluctuer entre les périodes glaciaires et interglaciaires du Pléistocène. Pendant près de 2 millions d'années depuis le début du Pléistocène, le niveau des mers a fluctué dans une amplitude de 40 à 60 m sous les niveaux actuels sur des périodes de 41000 ans, donnant des masses continentales beaucoup plus grandes et une moindre proportion de mers peu profondes pendant les périodes glaciaire. Ainsi, un climat plus frais et plus sec a persisté dans le Sundaland avec de vastes basses terres à l'intérieur, sans doute sous la forme de couloirs de savane ou d'autres types d'écosystèmes de plaine et des poches de forêts tropicales à feuilles persistantes dans ses hautes terres extérieures. Pour les derniers ~ 800000 ans, le Sundaland a été confronté à une périodicité plus longue des cycles globaux glaciaires-interglaciaires de 100000 ans et

à des amplitudes plus larges des variations du niveau de la mer, qui ont atteint 120 m sous le niveau actuel. Pourtant, le niveau de la mer peut également augmenter comme à son niveau contemporain pendant la période interglaciaire, donnant un climat plus humide mais des masses terrestres beaucoup plus petites et fragmentées, correspondant à peu près à la moitié de son extension maximale.

Des études récentes suggèrent que les cycles de fluctuations du niveau de la mer ainsi que les changements climatiques et de connectivité terrestre ont influencé la diversification de la biodiversité du Sundaland. En modifiant les barrières au flux génétique, ces variations seraient à l'origine d'événements de vicariance et de dispersion tout au long du Pléistocène region. Confinée dans les masses continentales, l'évolution des organismes d'eau douce peut être affectée par un niveau de la mer plus bas pendant la période glaciaire, créant davantage de connexion des bassins versants d'eau douce du Sundaland (systèmes des paléorivières), facilitant ainsi la dispersion par rapport aux bassins versants fragmentés des périodes interglaciaires. Quatre de ces systèmes de paléorivière ont été identifié dans le Sundaland, à savoir les paléorivières de la Sonde Est, de la Sonde Nord, du détroit de Malacca et la paléorivière du Siam. Ces paléorivières trouvaient leurs sources dans les masses continentales émergentes et s'écoulaient principalement dans les zones de plaines de l'intérieur du Sundaland avant d'atteindre la mer.

Ce travail de thèse a abordé cette question en utilisant des approches moléculaires basées sur la coalescence pour explorer les patrons spatiaux et temporels au cours de la diversification des espèces de poissons d'eau douce de Sundaland. Une attention particulière a été donnée aux hypothèses de pompe à espèces du Pléistocène et des Paléorivières. Selon la première hypothèse, les cycles de connexion et de déconnexion des masses terrestres du Sundaland créés par les fluctuations du niveau de la mer au Pléistocène auraient augmenté les chances de colonisation et augmenté les taux de spéciation, générant ainsi une quantité extraordinaire de biodiversité au Sundaland. En particulier pour les biotas d'eau douce du Sundaland, «l'hypothèse de la pompe à espèces du Pléistocène» a toujours été étroitement liée aux paléorivières. L'hypothèse des paléorivières prédit que les fluctuations du niveau de la mer du Pléistocène, ayant abouties à des changements cycliques de connectivité des milieux aquatiques, ont aboutit à un taux de diversification plus élevé des biotas d'eau douce.

Des approches moléculaires basées sur la coalescence ont été utilisées au lieu de la taxonomie morphologique traditionnelle, et ce afin d'éviter des biais en regroupant différentes lignées sans différences morphologiques apparentes (diversité cryptique), bien qu'elles puissent également être distinctes évolutivement. Surtout dans le Sundaland, de nombreuses études ont suggéré l'abondance de ces lignées cryptiques chez les poissons d'eau douce. Ainsi, la dépendance envers les espèces morphologiques,, utilisées par la taxonomie traditionnelle, dans l'étude phylogéographique du Sundaland peut entraîner une surestimation du délai de spéciation et / ou une utilisation invalide de la distribution des taxons. Dans cette étude, des approches de délimitation génétique des espèces ont donc été utilisées pour définir des Unités Taxonomiques Moléculaires Opérationelles (Molecular Operational Taxonomic Units – MOTU). Plusieurs méthodes de délimitation des espèces ont été utilisées, mais chacune présentant des limites, un consensus entre les différentes méthodes a été établit afin de produire une délimitation de délimitation robuste des espèces. Ainsi, nous avons utilisé quatre méthodes de délimitation, dont une méthode basée sur les distances génétiques avec Automatic Barcode Gap Discovery (ABGD), une méthode basée sur les réseaux avec Refined Single Linkage (RESL), et deux méthodes basées sur les arbres phylogénétiques, la méthode utilisant une distribution de poisson des évènements de spéciation et mutation (PTP) et le Generalized Mixed Yule Coalescent (GMYC). Un schéma de délimitation final a ensuite été établi sur la base d'un consensus de 50%.

Pour soutenir notre objectif général, nous avons d'abord cherché à évaluer l'adéquation entre la distribution des lignées moléculaires de plusieurs taxons avec les limites des bassins versants des paléorivières, en compilant des données de séquences et de distribution de la littérature. Dans cette première étude, nous avons utilisé les données moléculaires disponibles dans les bases de données internationales pour des familles avec une bonne couverture taxonomique et spatiale, représentative en Asie du Sud-Est (et plus particulièrement dans le Sundaland), pour répondre aux questions suivantes: 1) les paléorivières servaient-elles de couloirs de dispersion entre les îles pendant les périodes glaciaires du Pléistocène ?; 2) les bassins versants des paléorivières ont-ils initié une divergence allopatrique à travers leurs frontières ?; et 3) les fluctuations climatiques du Pléistocène ont-elles augmenté les taux de diversification des espèces. Des phylogénies calibrées temporellement à partir d'espèces délimitées par l'ADN ont été déduites à partir de 2211 séquences de

1511 individus, représentant 110 espèces nominales de six genres de poissons d'eau douce riches en espèces en Asie du Sud-Est (*Clarias*, *Channa*, *Glyptothorax* et Zenarchopteridae). Les résultats mettent en évidence un niveau élevé de diversité cryptique avec un nombre total de MOTU pour *Clarias*, *Channa*, *Glyptothorax* et Zenarchopteridae de 29, 120, 61, 43, respectivement; alors que le nombre de leurs espèces nominales est respectivement de 13, 30, 50, 17. Plus de la moitié des espèces nominales et 88% des MOTU trouvent leurs origines au Pléistocène. Malgré cela, les analyses de diversification indiquent que les modèles de diversification dépendant du niveau de la mer rendent mal compte des dynamiques de prolifération des espèces pour tous les clades, à l'exception du genre *Channa*. Les genres *Clarias* et *Glyptothorax* présentent des taux de spéciation constanst tandis que les Zenarchopteridae présente un profil de diversification croissante au fil du temps. Bornéo est identifié comme l'origine la plus probable pour la plupart des lignées de Sundaland, avec deux vagues de dispersion vers Sumatra et Java au cours des 5 derniers millions d'années. Environ 60% des événements de spéciation sont associés aux limites des paléorivières, plus souvent qu'aux limites des îles (40%). Au total, un tiers des événements de spéciation se sont produits au sein des paléorivières au sein des îles, ce qui suggère que l'hétérogénéité de l'habitat et des facteurs autres que l'allopatrie entre les îles ont considérablement affecté la diversification des poissons du Sundaland. Ces résultats sont traités dans le **Chapitre 1**, intitulé « Impact des fluctuations eustatiques du Pléistocène sur les dynamiques évolutives dans les points chauds de la biodiversité de l'Asie du Sud-Est ».

En utilisant les résultats de la première étude comme référence, nous avons ensuite appliqué les mêmes questions de recherche pour estimer l'âge des clades et la répartition géographique des lignées de la sous-famille des Rasborinae et examiner leur adéquation avec les attendus de l'hypothèse des paléorivières. Pour cela, nous avons utilisé un ensemble de données empiriques nouvellement générées chez les Rasborinae, un groupe répandu et extrêmement diversifié de poissons d'eau douce primaires de Sundaland, qui a été souvent considéré comme un groupe modèle pour exploree la biogéographie de la région. Malheureusement, les Rasborinae constituent un groupe problématique, en attente d'une révision taxonomique avant de pouvoir être analysé phylogéographiquement. Cette première étape est présentée au **Chapitre 2**, intitulé « Démêler la taxonomie de la sous-famille des Rasborinae (Cypriniformes, Danionidae) dans le Sundaland à l'aide des codes-barres ADN ». Au cours de cette

étude, une bibliothèque de référence de codes-barres ADN des Rasborinae de Sundaland a été créée pour examiner les contours des espèces et leurs répartitions spatiales grâce à des méthodes de délimitation des espèces basées sur l'ADN. Une liste de référence des Rasborinae de Sundaland a été compilée à partir de catalogues en ligne et utilisée pour estimer la couverture taxinomique de l'étude. Un total de 991 codes-barres ADN provenant de 189 sites d'échantillonnage dans le Sundaland ont été générés et compilés avec 106 séquences précédemment publiées, assemblant une bibliothèque de référence de 1097 séquences qui couvre 65 taxons, dont 61 des 79 espèces connues de Rasborinae de Sundaland. La délimitation génétique des espèces a détecté 166 MOTU distincts, identifiés par l'absence de chevauchement entre les distributions des distances génétiques intraspécifiques et interspécifiques.

Ces MOTU ont ensuite servi de base à la phylogéographie des Rasborinae présentée au **Chapitre 3**, et intitulée « Dispersion limitée et diversification in situ déterminent l'histoire évolutive des poissons Rasborinae dans le Sundaland ». Dans ce travail, en plus d'utiliser l'ensemble de données de codes-barres ADN et les MOTU résultants de travaux présentés au Chapitre 2, nous avons également produit 58 génomes mitochondriaux complets (mitogénomes). Des mitogénomes complets supplémentaires provenant de bases de données internationales ont été ajouté pour reconstruire l'arbre des Rasborinae par une approche de phylogénomique. Cet arbre a ensuite servit de base pour tester diverses prédictions de l'hypothèse des paléorivières. Au total, nous avons agrégé 1017 codes-barres ADN et 79 mitogénomes. Les estimations des aires ancestrales des espèces ont été réalisées en utilisant à la fois un partitionnement en île et en paléorivières pour examiner l'impact de la connectivité insulaire pendant l'eustasie du Pléistocène sur la dispersion. Les tendances temporelles de la diversification ont été explorées à l'aide d'approches probabilistes. Les résultats ont montré que les Rasborinae de Sundaland était originaire d'Asie continentale, une origine daté à 33,41 Ma. Quatre clades majeurs sont identifiés avec des ancêtres dont les âges sont estimés entre 31,10 et 25,95 Ma. Bornéo et la paléorivière du North Sunda sont identifiés comme centres d'origine. Les modèles géographiques de spéciation indiquent que la plupart des événements de spéciation se sont produits dans les îles et que les modèles à taux de spéciation constants sont les plus probables pour tous les clades.

Dans l'ensemble, ces travaux ont détecté un niveau élevé de diversité cryptique, avec des ratios de MOTU / espèces nominales de 1,83, 2,00, 2,00, 2,23 , 3.17 et 3.6

pour *Clarias*, *Dermogenys*, *Glyptothorax*, Rasborinae, *Hemirhamphodon* et *Channa*, respectivement. Considérant que Sundaland héberge environ 75% de l'ichtyodiversité d'eau douce indonésienne connue sur la base d'espèces nominales valides, cela montre l'importance de prendre en compte la diversité cryptique. La comptabilisation de la diversité cryptique peut augmenter massivement le niveau total attendu de diversité, un biais taxonomique majeur qui devra être prise en compte pour toute étude ultérieure et / ou applications en conservation.

Compte tenu de la proportion importante de diversité cryptique, les travaux de taxinomie moléculaire sont essentiels à l'étude de l'ichtyodiversité d'eau douce de Sundaland. Comme illustrée par cette étude, ces résultats valident un certain nombre de travaux antérieurs en confirmant: 1) la validité de certaines espèces récemment décrites (par exemple *Brevibora exilis* Liao & Tan, 2014); et 2) la monophylie de certains groupes (par exemple *Boraras* Kottelat et Vidthayanon, 1993). Une révision taxinomique est nécessaire dans de nombreux, en particulier: 1) lorsque les lignées cryptiques d'un taxon ne constituent pas groupe monophylétique (par exemple, le genre *Rasbora* (polyphylétique)); 2) lorsque les espèces nouvellement décrites contiennent encore une diversité cryptique (par exemple *Rasbora patrickyapi*, Tan, 2009 - 3 MOTU); ou lorsque les taxons supérieurs nouvellement décrits ne sont pas monophylétiques (par exemple les genres *Trigonopoma* Liao, Kullander & Fang 2010 (paraphylétique)).

Les reconstructions de zones ancestrales ont révélé que les lignées de poissons d'eau douce du Sundaland provenaient d'Asie continentale, avec une colonisation entamé à l'Oligocène, confirmant ainsi l'hypothèse de l'établissement pré-pléistocène. Ces lignées sont entrées dans le Sundaland principalement par la paléorivière de la Sonde Nord à Bornéo et se sont ensuite dispersées dans d'autres parties de Sundaland via une dispersion à longue distance, souvent suivie d'une diversification *in situ*. Cela suggère que la partie de Bornéo de la paléorivière de la Sonde Nord est le centre d'origine le plus probable de la diversité régionale des poissons d'eau douce. Java a été colonisée en dernier et a été soumise à un niveau élevé de fragmentation de l'habitat d'eau douce en raison de sa géomorphologie, laissant la place à une forte diversification *in situ*. Les branches de la paléorivière de la Sonde Est, la seul paléorivière qui traverse cette île, ont été coupées dans des bassins versants étroits et confinés, qui ont conduit à une diversification *in situ* des lignées de poissons d'eau douce. La reconstruction biogéographique de Sumatra

implique un scénario compliqué car l'île est ancienne (contrairement à Java), présente une activité volcanique (contrairement à Bornéo) et est traversée par trois paléorivières. Généralement, la plupart des lignées de Sumatra ont colonisé l'île par les paléorivières de la Sonde Nord et de Malacca. Les lignées les plus anciennes enregistrées chez les Rasborinae (de la fin de l'Oligocène au début du Miocène) sont associées à ces systèmes de paléorivières et contrastent avec la plupart des lignées, qui sont plus jeunes. Ce schéma suggère de multiples événements de colonisation soit par le biais de la partie de Bornéo de la Sonde Nord ou directement du continent via la paléorivière du détroit de Malacca. Alors que la diversification *in situ* est, à nouveau, révélée chez de nombreuses lignées de Sumatra, cette île présente également plusieurs zones géographiques avec une affinité biogéographique incertaine probablement en raison de l'existence de paléorivières voisines circulant sur les mêmes basses terres, ce qui donne plus de chance de flux génétique.

Malgré la concordance dans la composition des espèces au sein d'une même paléorivière (dans la même île ou des îles différentes) et l'implication apparente de la dispersion, d'autres résultats remettent en question l'impact des fluctuations eustatiques du Pléistocène sur la dispersion des poissons d'eau douce du Sundaland. Nous avons constaté que, bien que l'abaissement du niveau de la mer pendant les périodes glaciaires reconnecte les bassins versants dans les paléorivières, cela n'ouvre pas nécessairement des canaux de dispersion inter-îles pour les poissons d'eau douce en raison du couloir de savane et d'écosystèmes forestiers saisonniers traversant l'intérieur du Sundaland et a servi de barrière à la dispersion en raison du manque d'habitat convenable. D'autre part, la perméabilité des limites physiques des bassins versants des paléorivières et la variabilité au sein des paléorivières ont créé respectivement un flux de gènes entre paléorivières en période interglaciaire et l'allopatrie au sein du paléorivière en période glaciaire. En raison de la géomorphologie complexe de la région, les limies entre paléorivières voisines peuvent être poreuses, par exemple, lorsqu'elle couvre des habitats d'eau douce de plaine généralement plats (par exemple entre les paléorivières de la Sonde Nord et de la Sonde Est dans le sud de Sumatra), générant ainsi plus de chances de flux de gènes. La géomorphologie peut également générer une spéciation allopatrique au sein des paléorivières, comme en témoigne le contexte géologique complexe de Java qui a fragmenté la paléorivière de la Sonde Est et conduit des radiations locales. Pendant ce temps, la variabilité de l'habitat peut déclencher une spéciation allopatrique au sein des paléorivières comme

de nombreux taxons d'eau douce du Sundaland le montre pour certains types d'habitats d'eau douce, créant une barrière pour le flux génétique même si ces habitats sont connectés au sein de la même paléorivière.

Sur le plan temporel, les fluctuations eustatiques du Pléistocène et la dynamique des paléorivières n'ont pas élevé le taux de diversification des poissons d'eau douce du Sundaland comme suggéré, à l'exception du genre *Channa* qui est opportuniste et présente une forte capacité de dispersion. Par ailleurs, aucun des taxons examinés, et présentant une proportion substantielle de MOTU datant du Pléistocène, ne présente de taux de diversification en baisse comme le suggère le modèle de diversification dépendante de la diversité (DDD), la plupart des groupes suivant un modèle à taux constant de spéciation. Dans ce cas, plutôt que de résulter d'une spéciation élevée et limitée par la capacité de charge des milieux, une grande proportion de lignées du Pléistocène résulte d'une spéciation constante et continue dans le temps.

Différent des régions tempérées où les fluctuations climatiques du Pléistocène ont grandement affecté l'évolution de ses organismes, la coexistence des espèces en système tropical reposeraient davantage sur les interactions biotiques, favorisant une adaptation à différentes niches écologiques. Cette hypothèse est conforme à notre principale conclusion, selon laquelle les fluctuations eustatiques du Pléistocène ne sont pas le seul moteur de la diversification des poissons d'eau douce. En d'autres termes, la diversité des poissons d'eau douce de Sundaland est élevée, non pas à cause des mécanismes d'isolement et de reconnexion pendant le Pléistocène, mais plutôt à cause de la stabilité à long terme, tant géologiquement qu'écologiquement, facilitant une prolifération élevée et la persistance des espèces. À leur tour, les fluctuations eustatiques du Pléistocène et la dynamique des paléorivières ne sont que des éléments abiotiques qui interagissent avec d'autres aspects tels que: la géomorphologie locale, la variabilité des écosystèmes locaux / de l'habitat et les traits de vie des organismes. Étant donné que les bassins d'eau douce de Sundaland sont vastes et couvrent des types très hétérogènes de terrain physique et d'habitat, nous pouvons souligner que les facteurs écologiques ont peut-être joué un rôle plus important dans le maintien de cette diversité que prévu. Cependant, les fluctuations eustatiques du Pléistocène ont put être eu un impact sur les dynamiques évolutives, mais peut-être plus limité aux taxons avec des traits de vie particuliers, tels qu'une

adaptabilité et une capacité de dispersion élevée, et pouvant s'adapter à une grande diversité d’habitats écologiques.

General Introduction

Biodiversity

Biodiversity is one of the most widely used concepts in life sciences with quite broad range of definitions used to describe it. Generally, we can refer Biodiversity as a whole variability of all organisms that have ever lived at all levels of biological organization, with emphasize on the genetic, species and ecological diversity from terrestrial, marine, and other aquatic ecosystems and ecological complexes of which they are being part of (Gaston and Spicer 2004). While genetic diversity encompasses genetic coding system of organism (nucleotides, genes, chromosomes) as well as its intraspecific variations; species diversity involved mainly the variability among different species; and on the larger scale, ecological diversity, covers all ecological variability of the populations through niches and habitats, up to the level of biomes (Fig. 0.1).

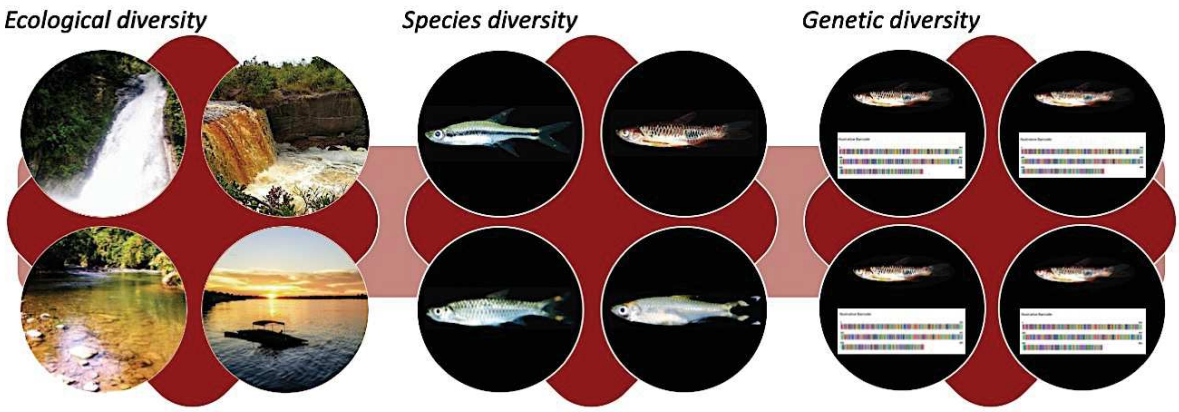

Figure 0.1 Levels of biodiversity

Given that these whole variabilities have been diversifying throughout history of life under influences from past and contemporary Earth's geological history as well as local ecological heterogeneity and dynamics (Avise 2000, 2009; Adamson et al. 2012), certainly Biodiversity has never been evenly distributed on Earth. Within the "Evolutionary Theatre", diversity of life comes and goes, diversifying through its birth (e.g. speciation, cladogenesis), death (e.g. extinction) and spatial movements. Chronologically, some groups which were pre-dominantly abundant for a certain duration could be highly scarce or even practically disappeared on other periods of

time; while spatial mapping of biodiversity during a same period of time also shows highly heterogenous pattern around the globe, both in their existence and abundance. Combining these aspects, Biogeography, especially Phylogeography, has been inquiring principles and processes governing spatial (geographic) distributions of genealogical lineages (micro and macroevolution) through time (Fig. 0.2) as depicted from its historical records on reconstructed phylogeny (Avise 2000).

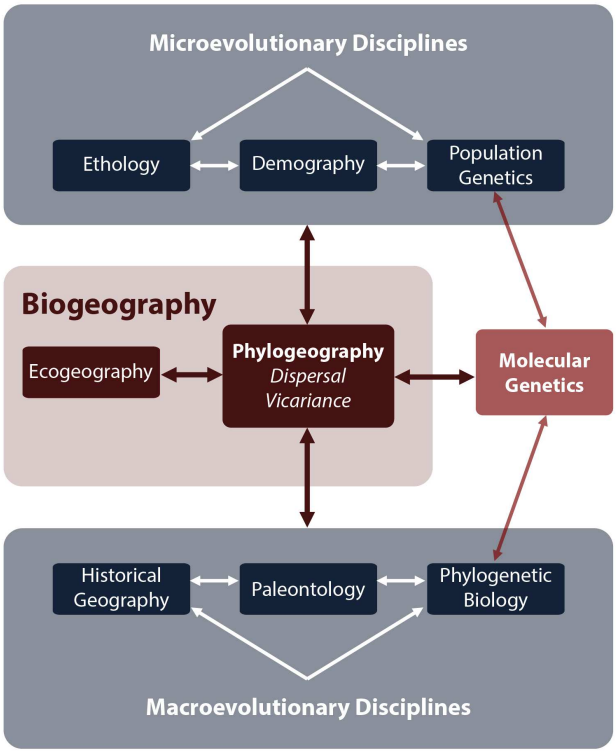

Figure 0.2 Framework of Phylogeography study (modified from Avise (2009))

Although naturally uneven, modern distribution pattern of global biodiversity shows even more extreme tendency of disparities with extensive and intensive human pressures have been creating significant changes in the existence of biodiversity. Within relatively short geological period of Anthropocene, human activities have endorsed biodiversity losses resulted in extinction of many life forms, dubbed as "the Sixth Mass Extinction" (Mittermeier et al. 2011), as well as cornering the remaining survivors into smaller and smaller pockets of environment. In the field of Conservation Biology, this phenomenon leads to the adoption of Biodiversity Hotspot terminology, which refers to the area featuring high level of endemic species and experiencing exceptional habitat loss due to anthropogenic pressure (Myers et al. 2000). While Biodiversity Hotspot is only one of many ways in explaining the state of global biodiversity, it is arguably one of the most popular approaches to use scientific data on biodiversity pattern in setting the scale of priorities on global conservation efforts for the last two decades. Using its standards, more and more areas have been given the status with current number hits 36 recognized hotspots, including four in Southeast Asia, namely: Indo-Burma, Philippines, Sundaland and Wallacea biodiversity hotspots (Figs. 0.3-0.4).

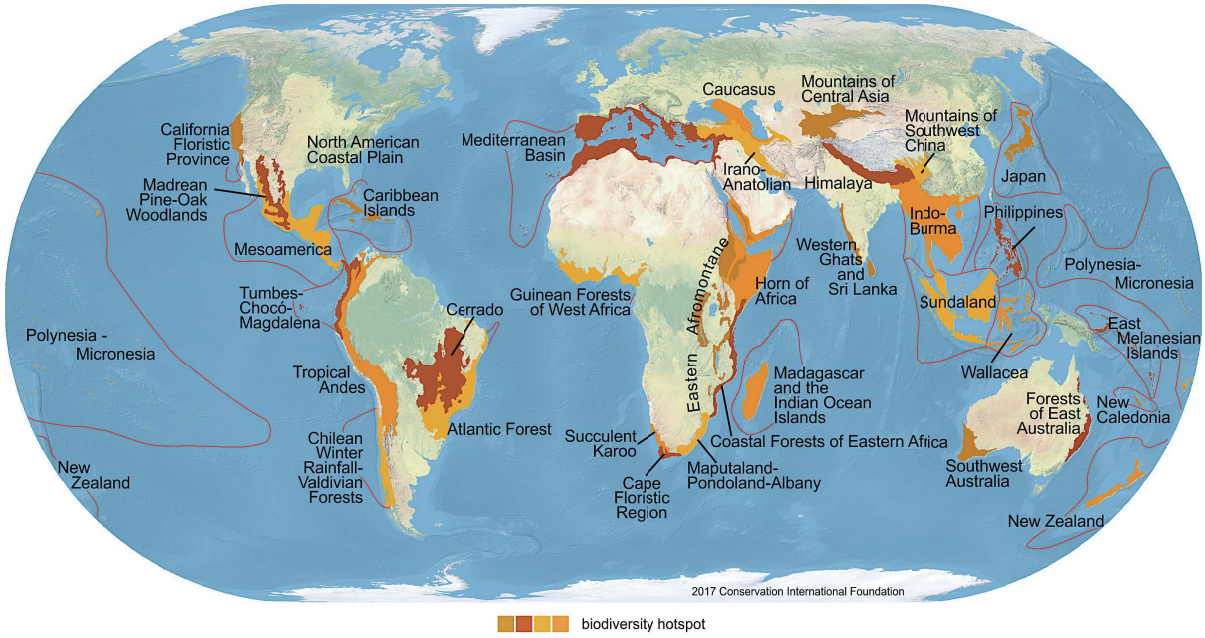

Figure 0.3 Current distribution of biodiversity hotspots
(Source: Conservation International, Wikimedia commons)

Sundaland

Sundaland is one of four biodiversity hotspots located in Southeast Asia, covering Malayan Peninsula, Sumatra, Borneo, Java and smaller islands around them (Lohman et al. 2011). Geologically sitting on the Sunda shelf, Sundaland (Fig. 0.4a) is separated by Kangar-Pattani Line and Huxley's Line on the north from (respectively) Indo-Burma and Philippines as well as by Wallace's Line on the east from Wallacea, while its west and south parts directly face Indian Ocean (Fig. 0.4). With equator cuts through its middle ground, all Sundaland enjoys warm tropical climate with heterogeneity within its parts can be attributed to the different bedrock materials, water drainages and topographies which comprise broad range of landscapes from flat plains on sea level to mountain ranges higher than 4,000 m above sea level. Considering this diversity, natural habitats in Sundaland covers many types of ecosystem, such as tropical forest (montane, lowland) and peat swamp forest in the terrestrial side; as well as various freshwater habitats like mountain torrents, streams, slow flowing rivers, lakes, swamps, lowland flood plains; brackish water habitat (e.g. mangrove forests); and shallow seas rich in marine biotas.

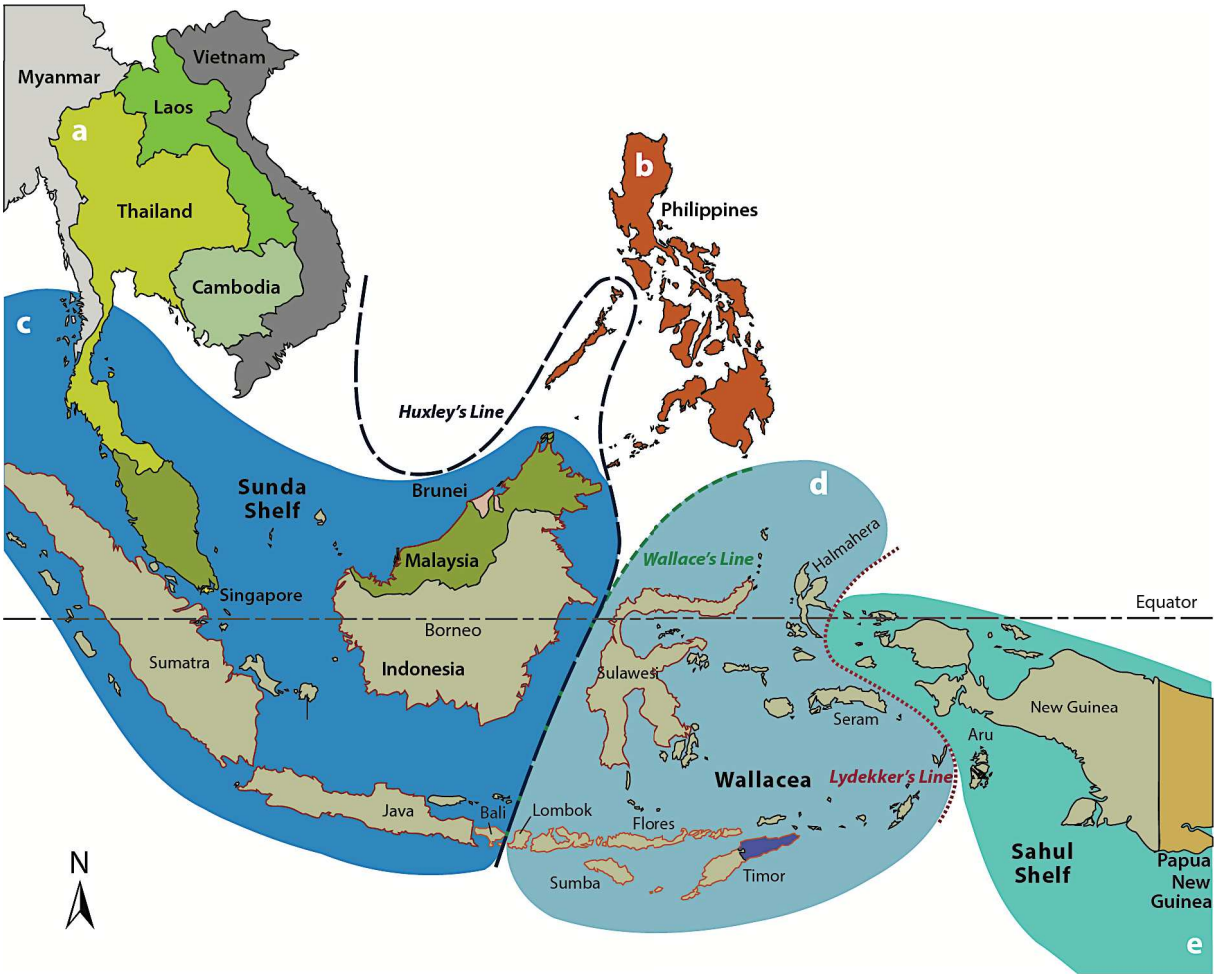

Figure 0.4 Map of Southeast Asia
(a) Indo-Burma biodiversity hotspot, (b) Philippines biodiversity hotspot, (c) Sundaland biodiversity hotspot, (d) Wallacea biodiversity hotspot, (e) Sahul shelf (modified from Lohman et al. (2011))

Biogeographically, biodiversity of Sundaland belongs to Oriental Zooregion and Palaeotropic-Malesian Phytoregion with its general pattern of biodiversity shows more resemblance towards mainland Southeast Asia compared with its neighbouring Wallacea and Papua (Sahul shelf). This region has high species richness and endemicity of higher plants, vertebrates, fish, amphibia, reptile, bird, and mammal (Myers et al. 2000) with recently updated data on freshwater fishes show even more remarkable record as even only Indonesian parts of insular Sundaland hosts nearly 900 of ca. 1,200 known Indonesian freshwater fish species which about 400 of them are endemics in the (Table 0.1). Coupled with high anthropogenic pressures from local and regional socioeconomic development, it puts Sundaland as one of the most threatened biodiversity hotspots on Earth. Furthermore, the importance of conservation efforts in the area is also signified by high occurrence of regional cryptic diversity, including on freshwater biotas (de Bruyn et al. 2004; Nguyen et al. 2008;

Pouyaud et al. 2009; de Bruyn et al. 2013; Hubert et al. 2015b; Dahruddin et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018; Hubert et al. 2019).

Table 0.1 Ichtyodiversity statistics of Indonesian parts of Insular Sundaland (Hubert et al. 2015b)

Island	Surface (km^2)	All Species				Endemics			
		N. of Family	N. of Species	Percent (Total All Species)	Density (Sp/1000 km^2)	N. of Species	Percent (Total All Endemics)	Percent (Total All Species)	Density (Sp/1000 km^2)
Bali	5,561	14	38	0.30	6.80	5	0.01	13.16	0.90
Bangka	11,330	23	35	0.90	3.10	10	0.02	28.57	0.88
Batam/Bintan	2,280	18	26	0.10	11.40	2	0.00	7.69	0.88
Belitung	4,800	14	18	0.20	3.80	4	0.01	22.22	0.83
Buru	9,505	14	16	0.20	1.70	0	0.00	0.00	0.00
Java	126,700	54	213	13.20	1.70	33	0.05	15.49	0.26
Kalimantan	539,500	66	646	35.30	1.20	294	0.46	45.51	0.54
Madura	5,290	11	12	0.30	2.30	1	0.00	8.33	0.19
Natuna	3,420	8	19	0.10	55.60	8	0.01	42.11	2.34
Sumatra	425,000	64	460	23.50	1.10	162	0.25	35.22	0.38
Total	1,133,386	73	899	74.00	0.80	431	0.68	47.94	0.38

Geological History of Sundaland

Biodiversity richness of Sundaland has been long subject on the study of Evolutionary Biology, especially by considering effects from Sundaland dynamic and complex geological history (Lohman et al. 2011). As can be seen from Figure 0.5, formation of Sundaland generally originated from Palaeocene (Fig. 0.5a) in early Cenozoic at ca. 60 millions of years ago (Ma) (Walker et al. 2018) when it grew as promontory at the southern tip of Eurasia. Later on, it underwent gradual changes following tectonic movements, combined along the way by the emergence of new islands in the area. Regional shallow seas around Sundaland started to open up in Late Miocene (10 Ma) and the contemporary configuration of Sundaland land masses has been generally stable since Early Pliocene (5 Ma).

Considering its dynamic geological history, it has been suggested that palaeoclimate in Sundaland also had changed through time as it was affected by the settlement and extent of Sundaland land masses and shallow seas. Generally, Sundaland palaeoclimate during Mio-Pliocene (20-5 Ma) was estimated to be warmer and wetter (pre-humid) (Lohman et al. 2011), while during Pleistocene, palaeoclimate could differ depending on the phase of global Pleistocene glacial cycles. Even though certainly not only affected Sundaland, impact of Pleistocene glacial cycles on the area was supposedly higher since Sundaland is surrounded by shallow seas which rarely

surpass 50 m in depth, hence the extent of its landmasses could change dramatically between glacial and interglacial periods as depicted in Figure 0.6.

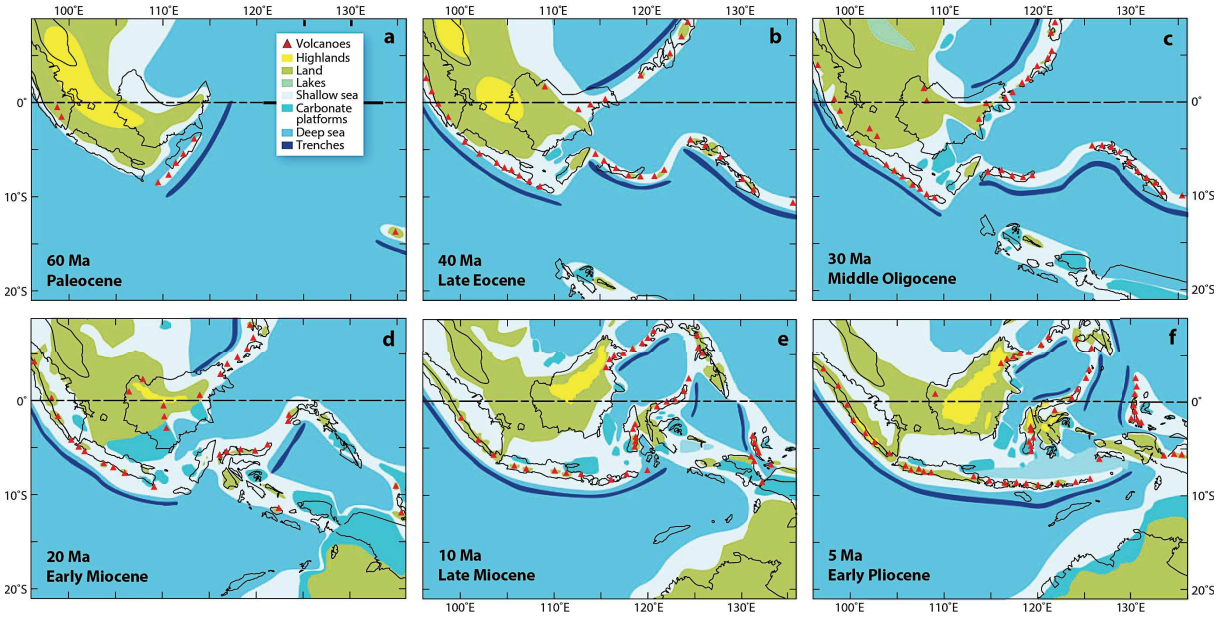

Figure 0.5 Reconstructed Cenozoic maps of Sundaland (modified from Lohman et al. (2011))

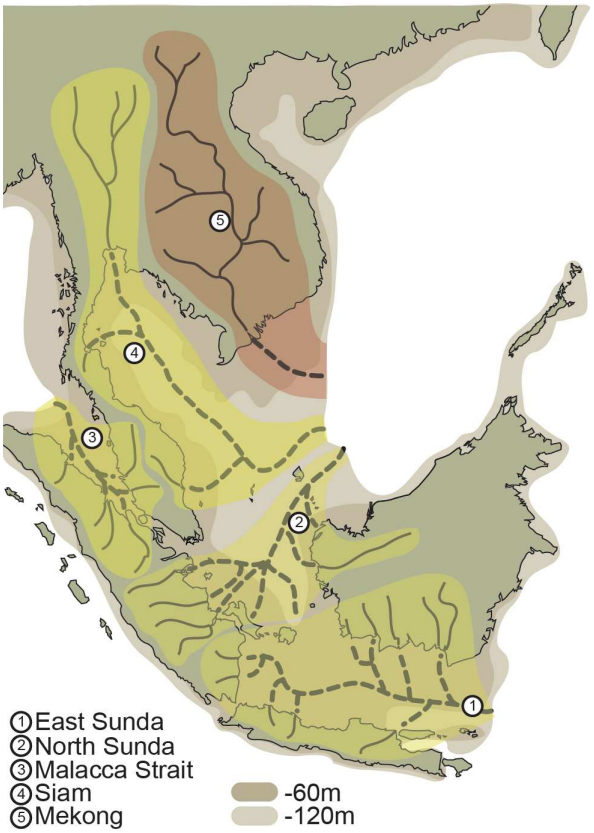

Figure 0.6 Maps of Sundaland palaeoriver systems

For almost 2 millions years since early Pleistocene, global sea-level (Fig. 0.7) fluctuated in the amplitude of 40 m to 60 m below present-day levels on ca. 41 kyr period (Lisiecki & Raymo, 2005; Lohman et al., 2011; Miller et al., 2005), giving much larger landmasses and less proportion of shallow seas both during the glacial and interglacial periods (Cannon et al., 2009; de Bruyn et al., 2014) (Fig 0.6, -60 m) hence cooler and drier climate persisted around Sundaland with expansive lowland in its interiors, arguably in form of savanna corridors (Heaney 1991; Bird et al. 2005) or some other types of lowland ecosystems (Slik et al. 2011)

and pockets of tropical evergreen forests in its exterior highlands. For the last ~800 kiloyears (kyr) though, Sundaland faced longer period of global glacial-interglacial cycles of 100 kyr (Lisiecki and Raymo 2005; Miller et al. 2005) and wider amplitudes of sea-level stands which can reach 120 m below contemporary level during glacial low stand (Voris 2000) (Fig. 0.6, -120 m) giving the similar cool and dry climate as before; yet sea level can also heighten as in its contemporary level during interglacial period, giving wetter climate but much more smaller and fragmented refugial landmasses (Fig. 0.6, green areas) with just about half its maximum range (Lohman et al. 2011).

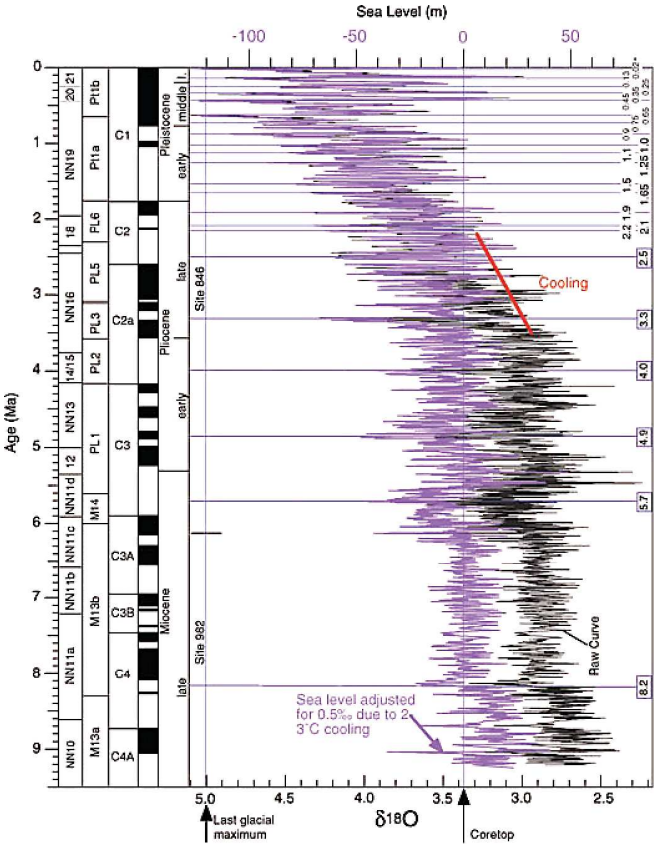

Figure 0.7 Global sea levels (Miller et al. 2005)

Recent studies have suggested that cycles of sea-level fluctuations as well as climatic and land connectivity changes have influenced diversification of Sundaland biodiversity by modifying barriers to gene flow hence the dynamics of vicariance and dispersal events throughout the Pleistocene evolutionary history in the region (de Bruyn et al. 2013; Hubert et al. 2015a). For terrestrial organisms, it means that during interglacial period, sea level rise created marine barrier (vicariance) as shallow water which fragmented smaller (refugial) landmasses lowered the opportunity of terrestrial dispersal. During glacial period, lowering sea level intensified and reconnected Sundaland landmasses hence promoting dispersal and gene flow for terrestrial organisms. For marine biota, the opposite happened with expanded land masses eventually become barrier (vicariance) for marine dispersal during glaciation and higher marine gene flow level happened during interglacial period.

Confined within landmasses, evolution of freshwater organisms followed the case of terrestrial biota with lower sea-level creating more interconnected freshwater

watersheds around Sundaland, usually called as palaeoriver systems (Fig. 0.6), that facilitated more dispersal compared with its currently fragmented freshwater basins (Voris 2000; Sathiamurthy and Voris 2006; Lohman et al. 2011). There were four recognized palaeoriver systems in Sundaland, namely East-Sunda (Fig. 0.6, number 1), North-Sunda (Fig. 0.6, number 2), Malacca Strait (Fig. 0.6, number 3), and Siam (Fig. 0.6, number 4) river systems. As can be observed from the figure, the palaeorivers were originated from the contemporary (different) emerging landmasses then mostly flowing through the lower altitude interior of Sundaland before ending in the sea. Based on this, we can expect that freshwater biota gene flows between currently different land masses were possible in the past, at least during Pleistocene glacial periods.

Diversification of Sundaland Freshwater Biota

High level of tropical biodiversity has often been dubbed as the result from the area serving as either "*museum*" or "*cradle*" for lifeforms (Moreau and Bell 2013). On the "*museum hypothesis*" point of view, the tropics (e.g. tropical rainforest) act as collections of old-persisting taxa with broad geographic ranges, relatively constant speciation rate but low extinction rate (Arita and Vázquez-Domínguez 2008); while on the "*cradle hypothesis*" point of view, the tropics are expected to harbour high speciation rate and low extiction rate (Stebbins 1974) thus serve as the "*centre of origin*" for species diversity (den Tex and Leonard 2013; Moreau and Bell 2013). In line with the latter, Sundaland biodiversity richness has often been attributed to the "*Pleistocene species pump hypothesis*" (Esselstyn et al. 2009; Brown et al. 2013; Papadopoulou and Knowles 2015a, 2015b; Li and Li 2018). Under this hypothesis, connection and disconnection cycles of Sundaland landmasses created by Pleistocene sea-level fluctuation has been thought to affect chances for dispersal and enhance speciation rate hence generate extraordinary amount of Sundaland biodiversity (den Tex et al. 2010).

For Sundaland freshwater biota, "*Pleistocene species pump hypothesis*" has always been closely connected to palaeoriver systems (Kottelat et al. 1993; Adamson et al. 2012), giving rise to "*Palaeoriver hypothesis*" in which Pleistocene sea-level fluctuation has been accredited with cyclical change of the extent and connectivity of, not only landmasses, but also freshwater watersheds of Sundaland and its biotas. Until Pliocene, both Sundaland land masses and freshwater basins were highly

interconnected (Fig 0.5), and only started to be fragmented as a consequence of the opening up of shallow seas around Sundaland in the early Pleistocene. Within that period of time, the highly interconnected Sundaland palaeoriver systems facilitated genetic homogenization by gene flows (Hubert et al. 2015a), hence dispersal of freshwater biota between Sundaland freshwater habitats, including between the currently fragmented parts due to its locations on the different contemporary islands. After that, cycles of glacial and interglacial periods and its attached sea-level fluctuations, with sea level drop during Pleistocene glaciation that could reach 120 m below present level, reconnected and re-disconnected Sundaland watersheds (Voris 2000; Sathiamurthy and Voris 2006; Lohman et al. 2011) which created such dynamics to promote higher level of freshwater biota diversification (higher diversification rate) hence the "*Pleistocene species pump hypothesis*".

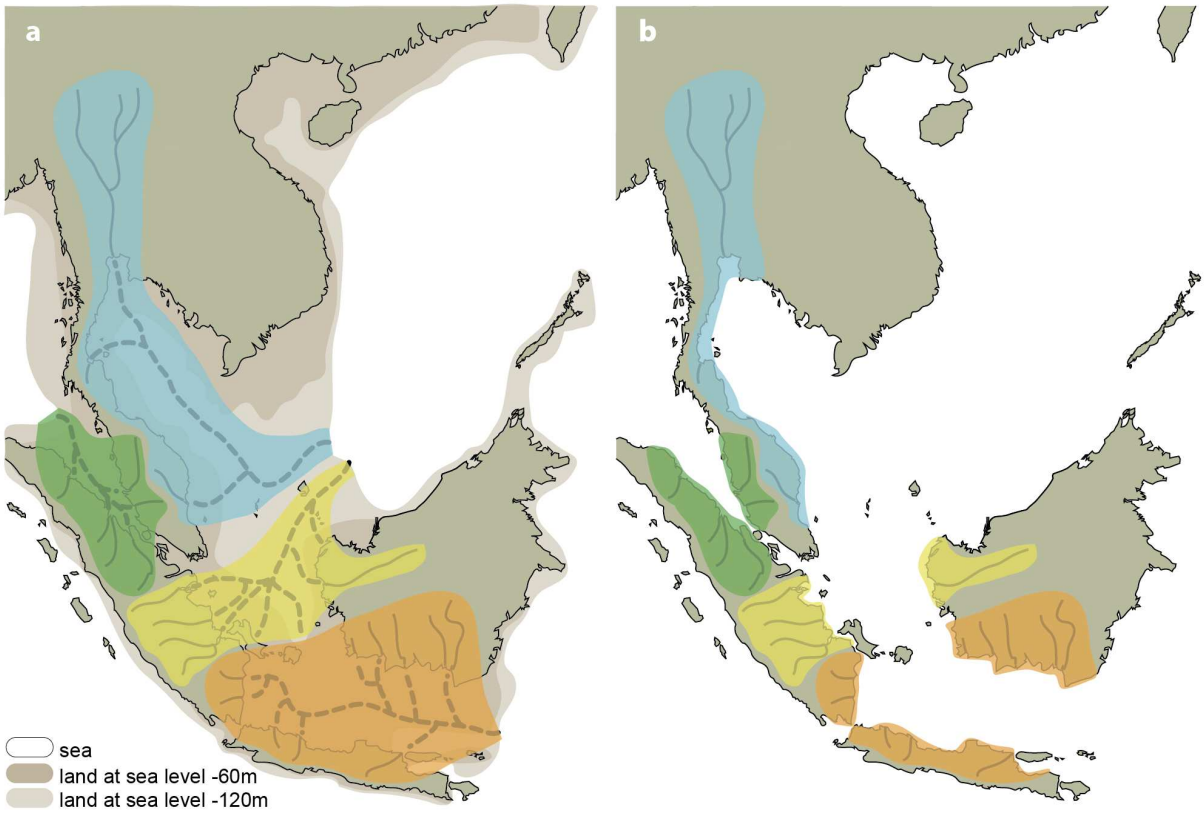

Figure 0.8 Sundaland freshwater basin connectivity change
(a) glacial connection and (b) interglacial fragmentation

Under "*Palaeoriver hypothesis*", during Pleistocene glacial periods, regional gene flows of freshwater biota including between the currently fragmented insular Sundaland were maintained through the existence of interconnected palaeoriver

systems (Kottelat et al. 1993; Voris 2000; Sathiamurthy and Voris 2006; Lohman et al. 2011; Hubert et al. 2015a) that acted as dispersal channels throughout the region (Fig. 0.8a). Meanwhile, during Pleistocene interglacial periods, especially for the last ~800 kiloyears (kyr) in which global glacial-interglacial cycles have longer periods and higher amplitudes of sea-level stand (Voris 2000; Lisiecki and Raymo 2005; Miller et al. 2005; Sathiamurthy and Voris 2006), fragmented freshwater basins were expected to serve as subjects of vicariance events with shallows seas around Sundaland submerged freshwater basins in the interior lowlands of Sundaland as well as blocked dispersal routes for freshwater biotas between the resulted fragmented refugial landmasses (Fig. 0.8b). The hindered dispersal was expected to endorse Pleistocene diversification of freshwater biota within, the also disconnected, remnant freshwater habitats.

On another point of view, it has been known that diversification through space and time are not only affected by plate tectonic movements or history of palaeoenvironment, but also their interplays with interaction between ecological dynamics of the biotas as well as landscape's physical and ecological properties (Hubert et al. 2015a). Using this understanding, diversification is seen as can be influenced by both regional spatial organization where dispersal and speciation take place as well as direct interaction among individuals on a local scale. The neutral geographically driven mechanism follows stochastic fluctuation of community's demography and genetic inflow from regional pool under the assumption that landscapes which have homogenous resources are hosting community with species that possess equivalent ecological functionality (Fig. 0.9a). On the other hand, direct local interactions among individuals is thought to give rise to biotic and abiotic constraints from ecosystems to community based on their respective ecological niches (ecological niche model), which can happen in form of habitat filtering by competition for resources (resource-dependent survival model, Fig. 0.9b) or life history trade-off mechanism (Fig. 0.9c). Putting this perspective into Sundaland, we can expect that its freshwater communities have been assembled historically from regional species pools after passing through abiotic (e.g. palaeoriver connectivity) as well as biotic ecological properties (e.g. niches, ecological competitiveness, life history traits) linked with them.

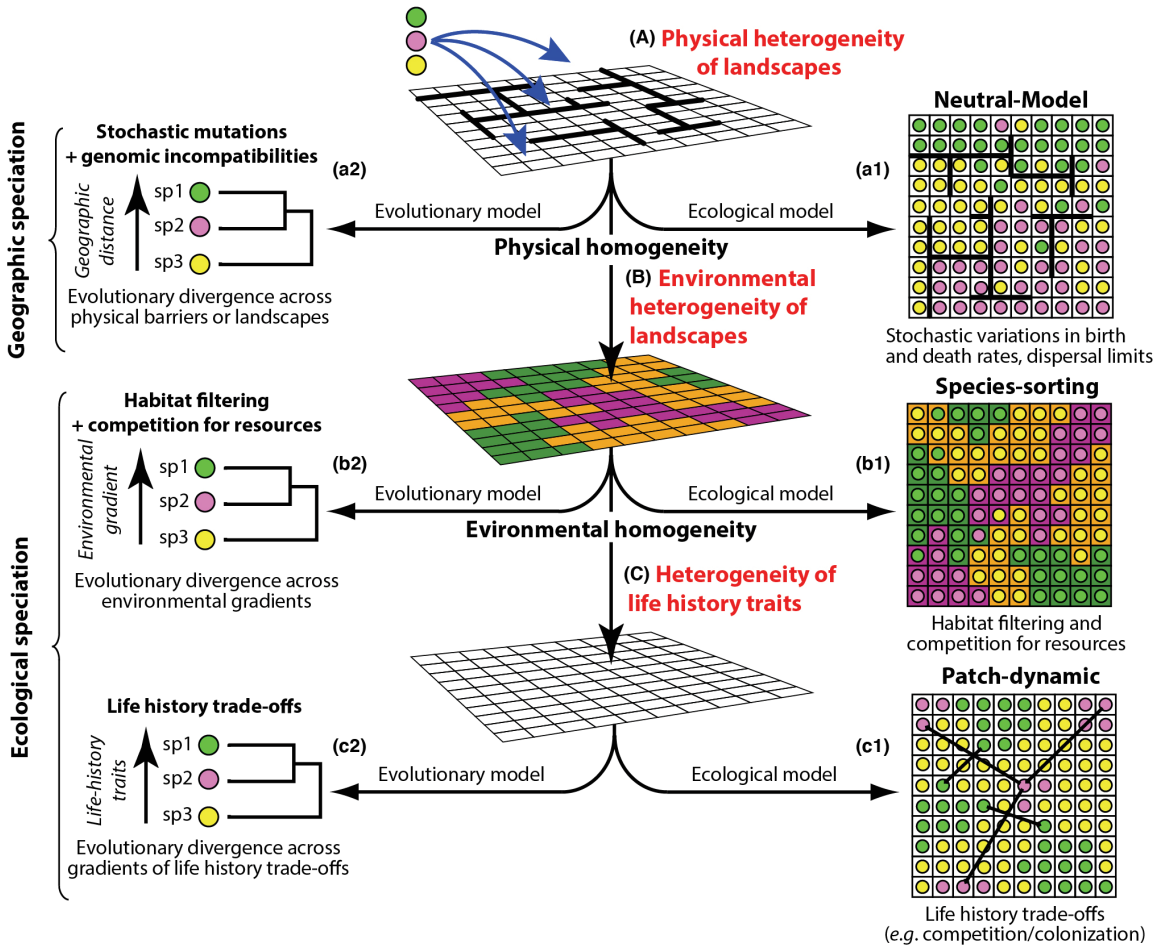

Figure 0.9 Speciation models
(modified from Hubert et al. (2015a))

Coalescent Theory and the Inventory of Biodiversity

Under traditional Taxonomy, even though molecular data may be involved, macroevolutionary aspect on Phylogeography study (Fig. 0.2) will normally use "*nominal species*" as unit of denomination for biodiversity in which evolutionary mechanisms are inferred during analyses. This has been broadly used since status of "*nominal species*" is seen as a more stable scientific consensus as well as the scientific data on biodiversity themselves have been historically recorded and managed mostly like this thus the widely and easily acquired information on it (Gaston and Spicer 2004).

Yet, this application has also been proven to come with various shortcomings. To begin with, "*nominal species*" status itself is hard to be defined discreetly since there are multiple species concepts that can be used in traditional Taxonomy, giving rise to multiple possible taxonomic status as well as possibility of inoperability of its use for certain group of organisms (Gaston and Spicer 2004). Moreover, it has also

been revealed that traditional Taxonomic practices in determining "*nominal species*" which have been heavily dependent to the phenotypic characteristics of organisms are prone to taxonomic bias by lumping together different lineages with no apparent morphological differences (cryptic diversity) despite they could also be evolutionary distinctive from each other (Hubert et al. 2015b, 2019). Especially in Sundaland, many studies have suggested the abundance of cryptic freshwater fish lineages (de Bruyn et al. 2013; Hubert et al. 2015a; Kusuma et al. 2016; Lim et al. 2016a; Beck et al. 2017; Dahruddin et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018) thus dependence towards "*nominal species*" in Sundaland Phylogeographic study may result in a bias, such as by overestimating timeframe of speciation and/or invalid use of reference distribution of taxon under examination.

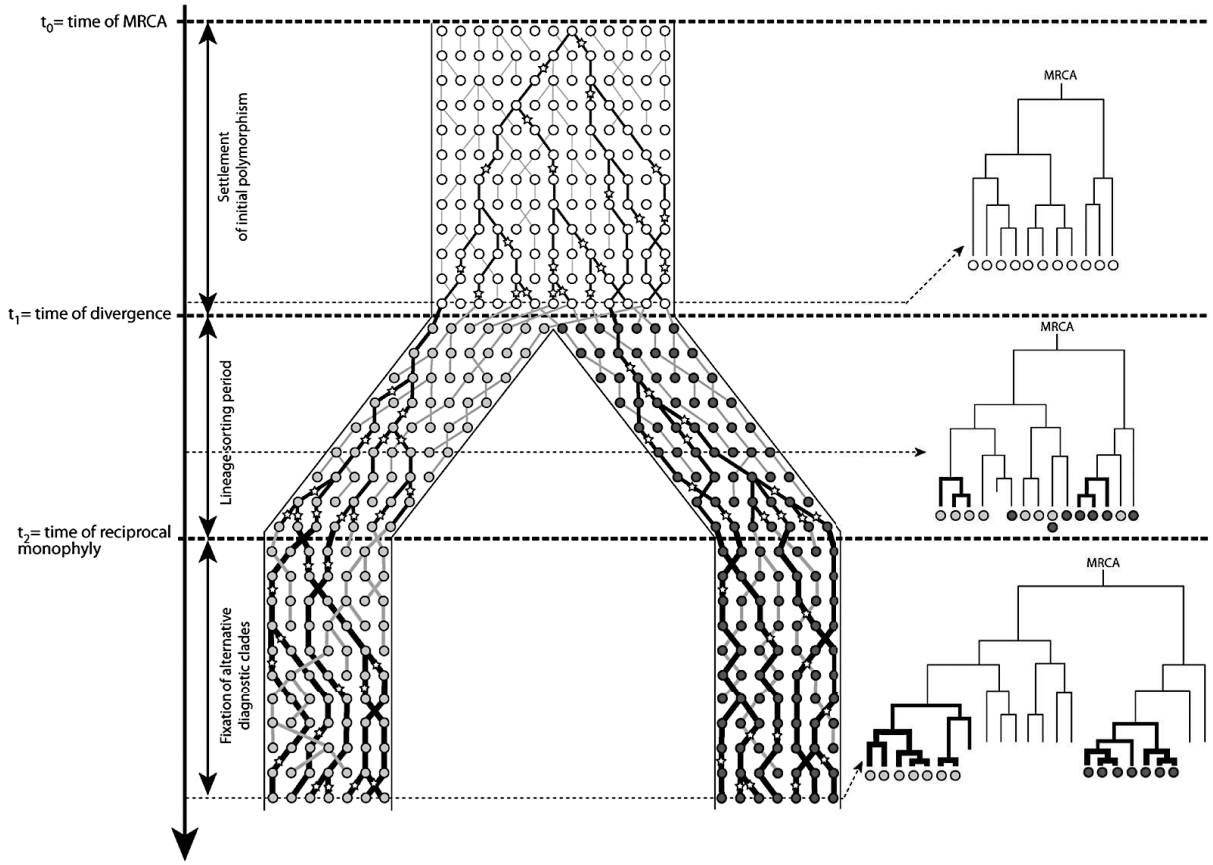

Figure 0.10 Coalescent theory
(modified from Hubert and Hanner (2015))

Reiterating framework on Figure 0.2, Phylogeography studies are concerned with the mechanisms governing spatiotemporal distributions of genealogical lineages during its evolution as depicted from phylogeny by utilizing molecular data alongside

tools from both macro and microevolutionary disciplines as well as ecogeographical information from the datasets (Avise 2000, 2009). In this sense, it should be clear that lineages examined in Phylogeography study should be evolutionary distinctive hence any possible bias from the still evolutionarily debatable status of "*nominal species*" in traditional Taxonomy is better to be avoided. To do that, rather than reconstructed based on traditional "*nominal species*" grouping, phylogeny used for Phylogenetic analysis would be better reconstructed using coalescent theory approach which is mathematical modelling of gene genealogies within and among related evolutionary distinctive species (Fig. 0.10; Avise 2000; Hubert and Hanner 2015).

Under coalescent theory, intraspecific genealogy (continuous microevolution) can be interpreted as a close-up observation of species level phylogeny (phylogenetic systematics/discreet macroevolution), in which by looking back in time, DNA haplotypes will eventually coalesce to the same common ancestor (Avise 2000; Hubert and Hanner 2015; Zachos 2018). On the opposite temporal direction though, we could also use coalescent theory to infer divergence mechanism of initially coherent population by using its genealogical intraspecific information (Fig. 0.10). With this approach, we can use molecular dataset from organisms to both: (1) estimate timeframe of divergence when population start to split genetically (Fig. 0.10, t_1), as well as (2) objectively delineate evolutionary distinctive lineages/evolutionary significant units (ESUs) (Hubert and Hanner 2015) when newly born genetically distinctive populations have reached their reciprocal monophyly (Fig. 0.10, t_2).

While in practical taxonomy the resulted ESU can be seen as an equivalent of "*species*" by applying "*phylogenetic species concept*" (PSC) hence the approach is dubbed as "*species delimitation method*" (SDM), more neutral term to be used will actually be "*operational taxonomic unit*" (OTU). One of the reasons for it is that there are different types of delimitation methods which can be categorized into at least three groups: (1) models defining branching patterns of phylogeny that rely more on on considering intraspecific coalescent dynamics (e.g. Automatic Barcode Gap Discovery (ABGD) and Refined Single Linkage (RESL)), (2) Kingman's model of gene sorting within species that relies more to interspecific branching patterns, and (3) mixed models which proportionally consider both the intraspecific (coalescent) and interspecific (phylogenetic) components (e.g. Generalized Mixed Yule Coalescent (GMYC)) (Hubert and Hanner 2015). Different assumptions used for each approach can be or can be not producing concordant set of ESUs. Furthermore, even though a

consensus can be reached upon considering different models (Hutama et al. 2017; Hubert et al. 2019), the resulted phylogenetic grouping can still be different with the other taxonomic grouping from application of different species concepts (e.g. with "*biological species concept*" (BSC)). In this case then, we can see that the use of OTU terminology (or Molecular OTU/MOTU when we consider only the molecular characteristics) can be important to keep a neutral distance and objectivity of Phylogeography analysis from the more debatable notion of "*species*" itself.

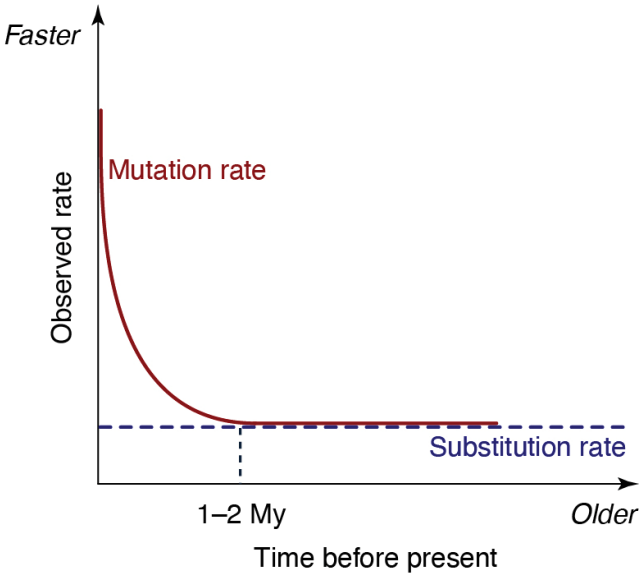

Figure 0.11 Molecular rate curve (modified from Ho and Larson (2006))

As has been mentioned before, coalescent theory can also be used to infer divergence time in each branching point along phylogeny by applying molecular clock approach which enables placement of independent time scales on evolutionary events (e.g. divergence events in Fig. 0.10, t_1) in form of "*mutation rate*" (instantaneous rate of nucleotide changes within genome) and "*substitution rate*" (rate of mutation fixation within population) (Fig. 0.11) (Ho and Larson 2006). In this research, we use specifically mixed model diversification/coalescent molecular dating approach by considering both intra and interspecific dynamics/variation of mutation accumulation in form of substitution rates.

Research Objective(s)

Despite there have been a lot of studies on Sundaland biodiversity, more scientific inquiries are still needed to explain what kind of evolutionary mechanisms are playing behind diversification of Sundaland aquatic biotas. In this study, our general objective is "*to explore time frame of vicariance and dispersal during diversity build-up of freshwater fish species in Sundaland*" by utilizing coalescence-based molecular

approaches. To support this investigation, we analyzed molecular dataset of Sundaland representative freshwater biota under two specific objectives as follow.

Since "*Pleistocene species pump hypothesis*" and "*Palaeoriver hypothesis*" have been considered as major drivers for diversification of Sundaland biodiversity, our first aim in this study is "*to assess the match between distribution of molecular lineages from multiple taxa with the palaeoriver boundaries*". This initial exploration used metadata analysis approach by utilizing the already existing molecular dataset with representative biological and spatial coverage in Southeast Asia (and specifically in Sundaland) which are *Clarias*, *Channa*, *Glyptothorax*, *Hemirhamphodon*, *Dermogenys*, *Nomorhamphus*. Results on this investigation will be covered in **Chapter 1**, entitled "*Impact of the Pleistocene Eustatic Fluctuations on Evolutionary Dynamics in Southeast Asian Biodiversity Hotspots*" which highlight significant level of cryptic diversity as well as extant MOTUs originated from Pleistocene speciation events despite predominant role of pre-Pleistocene geological settlement on biodiversity built-up and spatial arrangement as well as lack of dependence of biota diversification towards Pleistocene climatic fluctuations (PCF).

Second, at genus level, we aim "*to estimate clades' age and geographic distribution of* Rasbora *lineages*" in its relationship with Sundaland Pleistocene Palaeoriver Hypothesis. *Rasbora*, the widespread and extremely diversified group of primary freshwater fishes in Sundaland has been regarded as one of ideal candidates for phylogeographical study in the region (Liao et al. 2011). Unfortunately, *Rasbora* is also a problematic group which needs sufficient taxonomic clearance before it can be analyzed phylogeographically as has been done in the **Chapter 2**, entitled "*Disentangling the taxonomy of the subfamily Rasborinae (Cypriniformes, Danionidae) in Sundaland using DNA barcodes*" which generated a total of 991 DNA barcodes from 189 sampling sites in Sundaland, covering 61 of the 79 known Rasborinae species of Sundaland and resulted in 166 MOTUs based on the consensus of multiple delimitation methods. Next works in which we addressed a mitochondrial phylogenomic perspective of Sundaland Rasborinae diversification in its relationship with PCF and the underlining Palaeoriver hypothesis are covered in **Chapter 3**, entitled "*Limited dispersal and in situ diversification drive the evolutionary history of Rasborinae fishes in Sundaland*". In this chapter, on top of using COI dataset and the resulted MOTUs from Chapter 2, we also utilized 58 newly generated complete mitochondrial genomes of Sundaland Rasborinae to provide more robust phylogeny as backbone for the

subsequent phylogeography and diversification analyses as we have done in the first chapter.

Based on those inquiries, **Chapter 4**, entitled "*Synthesis on diversification of Sundaland aquatic biotas: build-up of freshwater fishes' diversity and distribution in a biodiversity hotspot*", will cover general discussion to give insight on spatiotemporal pattern of vicariance and dispersal during the accumulation of Ichtyodiversity in insular system of Sundaland as can be seen from the study. After that, implications on biological conservation efforts of Sundaland freshwater fishes within this biodiversity hotspot as well as the emerging perspectives based on this study will also be presented to sum up this study.

Chapter 1
Impact of the Pleistocene Eustatic Fluctuations on Evolutionary Dynamics in Southeast Asian Biodiversity Hotspots

Arni Sholihah[1,2], Erwan Delrieu-Trottin[2,3], Fabien L. Condamine[2], Daisy Wowor[4], Lukas Rüber[5,6], Laurent Pouyaud[2], Jean-François Agnèse[2], Nicolas Hubert[2]*

[1] Institut Teknologi Bandung, School of Life Sciences and Technology, Bandung, Indonesia.
[2] UMR 5554 ISEM (IRD, UM, CNRS, EPHE), Université de Montpellier, Place Eugène Bataillon, 34095, Montpellier cedex 05, France.
[3] Museum für Naturkunde, Leibniz-Institut für Evolutions und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Invalidenstrasse 43, Berlin, 10115, Germany.
[5] Division of Zoology, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Jalan Raya Jakarta Bogor KM 46, Cibinong, 16911, Indonesia.
[5] Naturhistorisches Museum Bern, Bernastrasse 15, Bern, 3005, Switzerland.
[6] Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, 3012, Bern, Switzerland.
*email: arni.sholihah@gmail.com

Systematics Biology; Accepted: 26 November 2020

Abstract

Pleistocene Climatic Fluctuations (PCF) are frequently highlighted as important evolutionary engines that triggered cycles of biome expansion and contraction. While there is ample evidence of the impact of PCF on biodiversity for continental biomes, the consequences in insular systems depend on the geology of the islands and the ecology of the taxa inhabiting them. The idiosyncratic aspects of insular systems are exemplified by the islands of the Sunda Shelf in Southeast Asia (Sundaland), where PCF-induced eustatic fluctuations had complex interactions with the geology of the region, resulting in high species diversity and endemism. Emergent land in Southeast Asia varied drastically with sea level fluctuations during the Pleistocene. Climate-induced fluctuations in sea level caused temporary connections between insular and continental biodiversity hotspots in Southeast Asia. These exposed lands likely had freshwater drainage systems that extended between modern islands: the Palaeoriver Hypothesis. Built upon the assumption that aquatic organisms are among the most suitable models to trace ancient river boundaries and fluctuations of landmass coverage, the present study aims to examine the evolutionary consequences of PCF on the dispersal of freshwater biodiversity in Southeast Asia. Time-calibrated phylogenies of DNA-delimited species were inferred for six species-rich freshwater fish genera in Southeast Asia (*Clarias*, *Channa*, *Glyptothorax*, *Hemirhamphodon*, *Dermogenys*, *Nomorhamphus*). The results highlight rampant cryptic diversity and the temporal localization of most speciation events during the Pleistocene, with 88% of speciation events occurring during this period. Diversification analyses indicate that sea level-dependent diversification models poorly account for species proliferation patterns for all clades excepting Channa. Ancestral area estimations point to Borneo as the most likely origin for most lineages, with two waves of dispersal to Sumatra and Java during the last 5 Myrs. Speciation events are more frequently associated with boundaries of the palaeoriver watersheds, with 60%, than islands boundaries, with 40%. In total, one-third of speciation events are inferred to have occurred within

palaeorivers on a single island, suggesting that habitat heterogeneity and factors other than allopatry between islands substantially affected diversification of Sundaland fishes. Our results suggest that species proliferation in Sundaland is not wholly reliant on Pleistocene sea-level fluctuations isolating populations on different islands. [**Milankovitch cycles**; **eustatic fluctuations**; **insular systems**; **diversification**; **vicariance**; **dispersal**; **palaeoenvironments**; **freshwater fishes**]

Introduction

Pleistocene Climatic Fluctuations (PCF) are widely considered to be drivers of species diversification and distribution patterns (Dodson et al. 1995; Haffer 1997; Lohman et al. 2011; de Bruyn et al. 2014) that contributed to the development of highly heterogeneous patterns of species richness and endemism (Voris 2000; Nores 2004; Hubert and Renno 2006; Pellissier et al. 2014). Temporally located between 2.58-0.012 millions years ago (Ma) (Walker et al. 2018), PCFs result from the dynamic interactions between tectonic movements and oscillations of Earth's orbit, leading to fluctuations of average temperature and global climate. These Milankovitch cycles have impacted biomes (Gathorne-Hardy et al. 2002; Currie et al. 2004; Bird et al. 2005; Cannon et al. 2009) through palaeoenvironmental dynamics such as eustatic fluctuations. This caused physical fragmentation of terrestrial and freshwater biotas and shaped contemporary distribution patterns (Dodson et al. 1995; Gathorne-Hardy et al. 2002; Beck et al. 2017). Climate changes are frequently invoked to account for heterogeneous speciation and extinction patterns in temperate and tropical biomes. For instance, the Late Pleistocene Hypothesis pinpoints glacial cycles and associated sea-level fluctuations as the main drivers of increased speciation in the Northern Hemisphere (Barraclough and Nee 2001; Wiens and Donoghue 2004; Mittelbach et al. 2007; Beck et al. 2017). In the tropics, repeated marine incursions and lowlands drying resulted in the fragmentation of highlands (Haffer 1997; Nores 1999; Hubert and Renno 2006) and shrinking of terrestrial biotas into refugia, leading to extinction of coastal organisms (Condamine et al. 2015b) or the divergence of isolated populations across refugia (Bates et al. 1998; Nores 1999; Hubert and Renno 2006; Condamine et al. 2015). Pleistocene Climatic Fluctuations are frequently considered to be a “species pump” that offered increased opportunities of speciation through cycles of alternating species’ ranges contraction and populations’ fragmentation during glacial maxima with dispersal and species’ ranges expansion during inter-glacial times (Esselstyn and

Brown 2009; Brown et al. 2013; April et al. 2013; Papadopoulou and Knowles 2015a, 2015b; Li and Li 2018). As such, PCFs created opportunities of speciation in otherwise physically stable landscapes in the short-term. On the long-term,PCFs interacted with geology, and resulted in intricate outcomes. Some of these constitute model systems to explore the consequences of PCFs and associated sea-level fluctuations on spatial biodiversity patterns (Weigelt et al. 2016). The Pleistocene Aggregate Island Complex (PAIC) model, for instance, describes the species pump hypothesis on islands, most notably in the Philippines, Caribbean and Mediterranean archipelagos (Esselstyn and Brown 2009; Brown et al. 2013; Papadopoulou and Knowles 2015a, 2015b). Understanding the impact of PCFs on diversity patterns is particularly challenging for the implementation of science-based conservation in the most diverse and threatened biotas globally *i.e*. biodiversity hotspots (Myers et al. 2000; Hoffmann et al. 2010).

The biodiversity hotspots in Southeast Asia occur in an area where Pleistocene eustatic fluctuations interact with geology (Voris 2000; Woodruff 2010; Lohman et al. 2011). Sundaland (Java, Sumatra, Borneo and peninsular Malaysia) exemplifies a tropical insular region with complex interactions between climate and geology. Sundaland emerged during the early Cenozoic (ca. 66 Ma) as promontory at the southern end of Eurasia (Lohman et al. 2011). The isolation of Sundaland, however, is much more recent (Fig. 1.1). Complex tectonic movements and subduction activities during the Miocene triggered the formation of Borneo between 20 and 10 Ma (Figs. 1.1a & 1.1b) and the subsequent emergence of Sumatra and Java between 10 and 5 Ma (Figs. 1.1b & 1.1c). From a geological perspective, Sundaland's configuration has been stable since 5 Ma (Lohman et al. 2011; Hall 2013). However, insular Sundaland was not completely separated from mainland Southeast Asia until the Pliocene (5.33-2.58 Ma) (Hall 2009; Lohman et al. 2011). Upon entering the early Quaternary (2.58 Ma), sea level dropped from 40 to 60 m below present levels over a ca. 41 kyr period (Lisiecki and Raymo 2005; Miller et al. 2005; Cannon et al. 2009; Lohman et al. 2011; de Bruyn et al. 2014) forming land bridges among the Sunda islands and mainland (Fig. 1.1d, -60 m). Around 800 kyr, Sundaland experienced longer glacial cycles of 100 kyr and lower sea level during eustasy (Lisiecki and Raymo 2005; Miller et al. 2005) which enlarged land bridges and increased fragmentation during interglacial times. During the Last Glacial Maximum (LGM) ca. 17 kyr, sea level dropped down to 120 m below its present level (Voris 2000; Sathiamurthy and Voris 2006) (Fig. 1.1d), resulting in the widest landmass extension to date. At the end of the LGM, rising sea levels

shrank terrestrial habitats, and Sundaland biotas are currently considered to be refugial (Cannon et al. 2009; Woodruff 2010; Lohman et al. 2011). Interactions between the geological history of subduction and PCFs are expected to have significantly affected the diversification of terrestrial and freshwater biodiversity on Sundaland (Voris 2000; Sathiamurthy and Voris 2006; Lohman et al. 2011; de Bruyn et al. 2013).

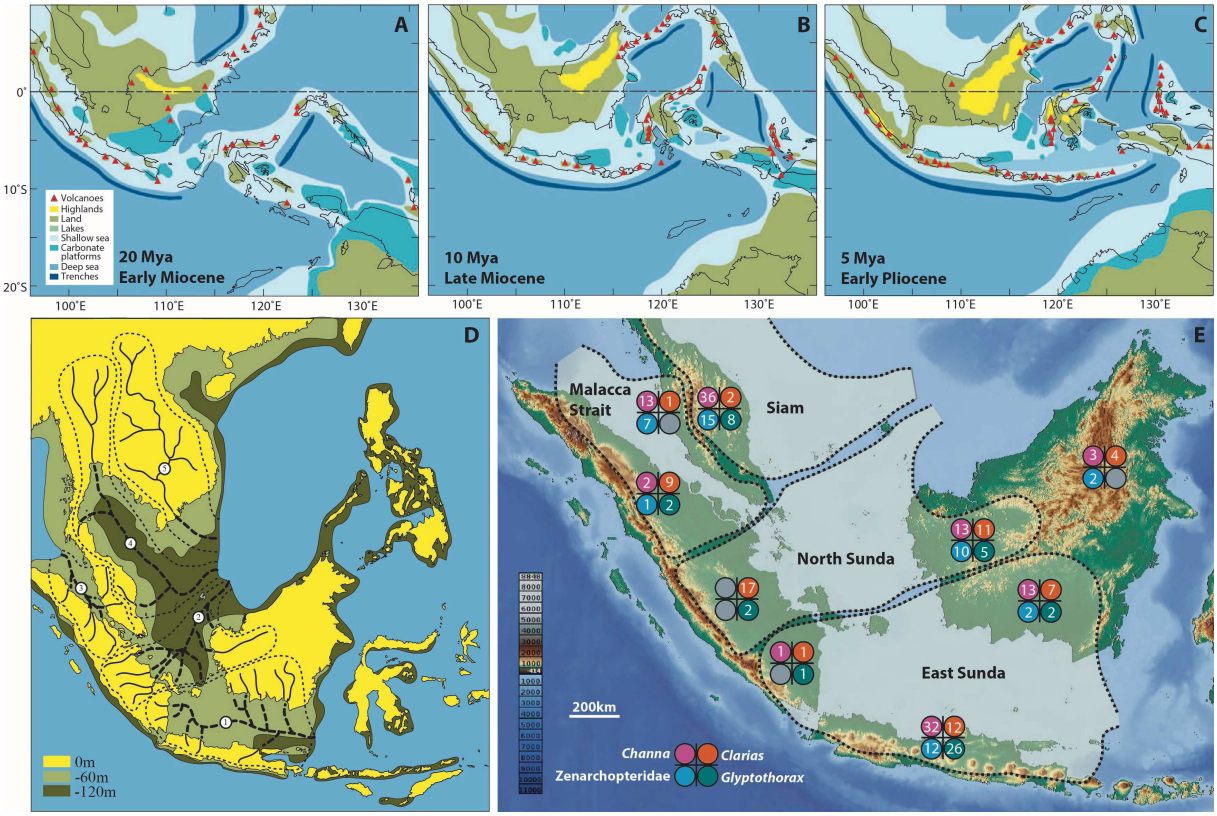

Figure 1.1 Palaeogeographic maps of Sundaland in the last 20 million years Pleistocene sea levels and associated palaeorivers

Reconstructions of historical land and sea distribution during the Neogene, depicting Sundaland during the early Miocene (A), late Miocene (B) and early Pliocene (C) (modified from Lohman et al. 2011). The Palaeoriver systems in Southeast Asia are shown (D), depicting East Sunda (1), North Sunda (2), Malacca strait (3), Siam (4) river systems of Sundaland and the Mekong river system of IndoBurma. Exposed lands during -60 and -120m sea level drops are illustrated in D. The cumulative number of sampling location for each group within the palaeoriver systems is provided (E) with colour codes as follows: orange for Clarias, green for Glyptothorax, light blue for Zenarchopteridae and rose for Channa, and grey circles indicate that no sample was obtained from those areas.

Eustatic fluctuations in Sundaland not only altered land exposure over time but also impacted the configuration of Sundaland's watersheds. During glacial maxima, sea levels were especially low, allowing riverine watershed to expand to newly exposed land that was previously sea floor (Voris 2000; Sathiamurthy and Voris 2006; Lohman et al. 2011). These temporary palaeoriver systems connected Sundaland's rivers and created temporary connections between Sundaland and continental

Southeast Asia. Particularly well-documented for the Pleistocene LGM, the lower sea level created four large palaeoriver systems that connected insular rivers across Sundaland's landmasses (Kottelat et al. 1993; Voris 2000; Woodruff 2010) including the palaeorivers of East Sunda (1 in Fig. 1.1d), North Sunda (2 in Fig. 1.1d), Malacca Straits (3 in Fig. 1.1d) and Siam (4 in Fig. 1.1d). For freshwater-dependent organisms, the existence of these palaeoriver systems has been long considered to be one of the main drivers of contemporary distribution patterns of freshwater organismal in the region (Kottelat et al. 1993; Voris 2000; Lohman et al. 2011). Formally known as the Palaeoriver Hypothesis, the impact of palaeorivers on freshwater species distribution has been investigated in *Hemibagrus nemurus* (Dodson et al. 1995), viviparous halfbeaks (de Bruyn et al. 2013), snakeheads (Tan et al. 2012), killifishes (Beck et al. 2017) and *Macrobrachium* shrimps (de Bruyn et al. 2004). These studies detected some congruence between lineages distribution and boundaries of palaeoriver watersheds. However, the determinants of these correlations are still debatable due to the limited number of taxa examined (Dodson et al. 1995; Lohman et al. 2011; Tan et al. 2012; de Bruyn et al. 2013; Beck et al. 2017). Furthermore, the ages of the taxa in these studies were generally older than the Pleistocene, casting doubt of the role of PCF in freshwater diversification (Dodson et al. 1995; Beck et al. 2017).

Here, we explore the evolutionary dynamics of species proliferation and spatial patterns of speciation among Southeast Asian freshwater fishes to address the following questions: (1) Did palaeorivers fragment populations of widely distributed taxa? (2) Did palaeorivers enable dispersal between islands during glacial maxima? (3) Did PCFs impact the pace of diversification?

From this set of initial questions, we derived several predictions made by the Palaeoriver Hypothesis regarding the evolution of species range boundaries during speciation (*e.g*. congruence between species range distribution and boundaries of palaeoriver watersheds) and the timing of speciation (*e.g*. Pleistocene species pump) to address the evolutionary response of Southeast Asian freshwater biotas to eustatic fluctuations. To test predictions made by the Palaeoriver Hypothesis, we compiled DNA sequences from widely distributed and well studied freshwater fish lineages with a variety of life history traits (Kottelat et al. 1993). Our molecular phylogenetic dataset of 1,511 individual was used to reconstruct the phylogenetic relationships and estimate the divergence times of 110 morphological species belonging to four of the most species-rich freshwater fish lineages in Southeast Asia: *Channa* (50 morphological

species), Zenarchopteridae (30 morphological species), *Glyptothorax* (17 morphological species), and *Clarias* (13 morphological species). Southeast Asian freshwater fishes are poorly known taxonomically and cryptic species are common (de Bruyn et al. 2013; Hubert et al. 2015a; Lim et al. 2016a; Beck et al. 2017; Nurul Farhana et al. 2018). Thus, all macroevolutionary inferences recognize Molecular Operational Taxonomic Units (MOTUs) delineated through DNA-based species delimitation methods (Blaxter et al. 2005; Ruane et al. 2014). Distribution ranges were characterized for these MOTUs and further used to explore spatial and temporal patterns of diversification through ancestral state reconstructions and birth-death modeling approaches.

Materials and Methods

Hypothesis Testing, Taxa Selection and Sampling

The Palaeoriver Hypothesis (Kottelat et al. 1993; Voris 2000; Woodruff 2010; de Bruyn et al. 2013) makes several predictions about the geography and timing of speciation. First, the palaeorivers might be expected to have promoted allopatric speciation among populations of species that were once widespread or expanding. If so, range boundaries between sister species that diverged during the Pleistocene might be expected to coincide with boundaries of the palaeoriver watersheds. Second, palaeorivers might have enabled dispersal between different islands, such as South Borneo and Java in the East-Sunda system during periods of low sea level (Fig. 1.1d). This hypothesis predicts that species assemblages will be most similar within palaeoriver drainage basins, even when those basins span two or more present-day islands. Assemblages between palaeoriver drainages on the same present-day island are likely to be markedly different in comparison. Finally, the interactions between geological history and PCFs have intensified during the last 5 Ma with the elevation of Sundaland predicting increased diversification. Estimating speciation events in different palaeorivers would allow testing of these hypotheses. However, different species interact with their environment in different ways, which will further influence their dispersal and colonization abilities. Thus, dispersal traits might be expected to influence the evolutionary responses of biotas to environmental changes (Barraclough and Nee 2001; Currie et al. 2004; McPeek 2008; Hubert et al. 2015a; Burbrink et al. 2016; Liu et al. 2019). To explore the three predictions made by the Palaeoriver

Hypothesis, we selected taxa conforming to the following requirements: (1) widespread distribution of lineages across Southeast Asia including Sundaland, (2) comprehensive sampling of described species within the taxon, (3) availability of nuclear and mitochondrial sequences, and (4) varying life history traits such as body size and dispersal ability.

Four lineages fulfilled these requirements: (1) the snakehead genus *Channa* (Anabantiformes, Channidae) widely distributed in Asia (Conte-Grand et al. 2017), (2) the walking catfish genus *Clarias* (Siluriformes, Clariidae) widely distributed in Asia (Pouyaud et al. 2009), (3) the catfish genus *Glyptothorax* (Siluriformes, Sisoridae) widely distributed in Asia (Kottelat et al. 1993; Jiang et al. 2011) and, (4) the halfbeak genera of Southeast Asia *Dermogenys*, *Hemirhamphodon* and *Nomorhamphus* (Beloniformes, Zenarchopteridae) (de Bruyn et al. 2013; Lim et al. 2016a; Nurul Farhana et al. 2018). Both *Clarias* and *Channa* can disperse easily. *Clarias* species, for instance, have a unique accessory air-breathing organ in the upper part of the branchial cavity that allows survival in oxygen-poor water and on land (Munshi 1961; Maina and Maloiy 1986). Meanwhile, *Channa* is a genus of air-breathing, predatory freshwater fishes with an extended swim bladder, paired supra-branchial chambers, and an interior labyrinth organ that enables them to live out of the water for several days. Several species can move on land and burrow into mud during droughts to keep their bodies wet (Kottelat et al. 1993; Berra 2001; Adamson et al. 2010; Serrao et al. 2014; Rüber et al. 2020). These capabilities make *Clarias* and *Channa* resilient to environmental variation, and several species are even considered invasive (Serrao et al. 2014; Conte-Grand et al. 2017). The other genera have lower dispersal abilities. *Glyptothorax*, the most species-rich genus of the catfish family Sisoridae, is smaller than *Clarias* and *Channa*, and is not equipped with an air-breathing organ (Berra 2001). *Glyptothorax* species are mostly confined, which fragment their distribution to isolated mountain rapid streams (Ng and Kottelat 2016; Hutama et al. 2017). The healfbeak genera are viviparous with limited larval dispersal abilities (de Bruyn et al. 2013; Nurul Farhana et al. 2018); although these genera are geographically widespread, their genetic lineages have more restricted distributions (Meisner 2001; Tan and Lim 2013; Lim et al. 2016a; Nurul Farhana et al. 2018). Our dataset comprised 2,211 sequences dataset from 1,511 individuals belonging to 110 nominal species and representing four lineages. Outgroups were then selected for each group, ranging from closely related to distantly related taxa following the phylogenetic classification of bony

fishes by Betancur-R et al. (Betancur-R et al. 2017). Outgroup selection was further refined according to previously published molecular studies for Clariidae (Pouyaud et al. 2009), Sisoridae (Jiang et al. 2011) and Channidae (Conte-Grand et al. 2017).

Genetic Species Delimitation

Genetic studies have identified substantial cryptic diversity in Southeast Asian freshwater fishes (Nguyen et al. 2008; Pouyaud et al. 2009; Hubert et al. 2015a; Dahruddin et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018; Hubert et al. 2019; Sholihah et al. 2020). Cryptic diversity could lead to biases in phylogenetic reconstruction and diversification analyses, particularly due to the overestimation of divergence age estimates in the absence of the closest relatives (Esselstyn et al. 2009; Patel et al. 2011; Ruane et al. 2014). To avoid these shortcomings, genetic species delimitation methods were used to define MOTUs. Several methods of species delineation have been developed, but each has pitfalls. Agreement among different methods suggest robust delimitation (Kekkonen and Hebert 2014; Kekkonen et al. 2015). Thus, we used four methods of species delimitation including a distance-based method with Automatic Barcode Gap Discovery (ABGD, Puillandre et al. 2012), a network-based method with Refined Single Linkage (RESL, Ratnasingham and Hebert 2013), and two tree-based methods including Poisson Tree Processes (PTP, Zhang et al. 2013) and the Generalized Mixed Yule Coalescent (GMYC, Pons et al. 2006). A final delimitation scheme was then established based on a 50% consensus (Hubert et al. 2019). All sequences used were aligned using MUSCLE (Edgar 2004) implemented in MEGA7 (Kumar et al. 2016) and manually edited.

These four delimitation methods use different input data. We used sequence alignments to carry out ABGD and RESL analyses. The ABGD analysis was performed through the online platform using the K2P substitution model (https://bioinfo.mnhn.fr/abi/public/abgd/abgdweb.html). The RESL algorithm was implemented through BOLD systems version 4 (http://www.boldsystems.org Ratnasingham and Hebert 2007), using Barcode Index Numbers (BIN, Ratnasingham and Hebert 2013). We used maximum likelihood (ML) phylogenetic trees for the PTP and GMYC analyses. For PTP, the analyses were carried out on the web service (https://mptp.h-its.org/#/tree). Only the single threshold version was used because preliminary analyses using the multiple threshold version of PTP were too conservative

compared to other methods (i.e. too few species). Maximum Likelihood phylogenies were reconstructed using RAxML (GUI 1.5) with the GTR+I+Γ model, and 5000 non-parametric bootstrap (BP) replicates were computed using RAxML-HPC Blackbox (Miller et al. 2010) with RAxML 8 (Stamatakis 2014). For GMYC, the multiple thresholds function (mGMYC) was implemented with the R-package SPLITS 1.0-19 (Fujisawa and Barraclough 2013). For GMYC, ultrametric trees were reconstructed using the chronopl function in the R-package *ape* 5.3 (Paradis et al. 2004; Paradis 2012; Paradis and Schliep 2019). Time calibration was constrained with the widely accepted fish substitution rate of 1.2% of genetic divergence per million years (Myrs) for mitochondrial protein-coding genes (Knowlton et al. 1992) and applied it to the maximum K2P distances, computed using mitochondrial protein coding-genes only, between descendant clades of selected nodes (Table 1.1). Reference nodes were selected to cover several levels of depth in the phylogenies and to cover a substantial portion of the total species diversity in the entire mitochondrial data set. A total of 13 nominal species were analyzed for *Clarias*, 50 nominal species for *Channa*, 30 nominal species for Zenarchopteridae and 17 nominal species for *Glyptothorax*.

Table 1.1 Reference nodes and associated time calibrations for each group
Node numbering is as in Figure 1.2. Age estimates result from the Bayesian StarBEAST2 analyses and parameter distributions represent median age, 5% and 95% quantiles for each reference nodes

Lineage	Calibrated node	Members	Age estimate - 1.2% clock rate (Ma)	Age estimate - StarBEAST2 (Ma)	Age estimate distribution - StarBEAST2
	Node 1 in Fig. 1.2a	Root	6.67 [4.00, 19.30]	7.09 [4.494, 10.310]	Normal (5% quantile 5.02, median 6.67, 95% quantile 8.31)
	Node 2 in Fig. 1.2a	*C. gariepinus* 1-2	1.25 [0.50, 3.00]	0.748 [0.252, 1.248]	Normal (5% quantile (-)0.395, median 1.25, 95% quantile 2.89)
Clarias (Clariidae, Siluriformes)	Node 3 in Fig. 1.2a	*C. batrachus 1-8*, *C. intermedius*, *C. kapuasensis*, *C. leiacanthus*, *C. macrocephalus* 1-4, *C. meladerma* 1-3, *C. microstomus*, *C. nieuhofii* 1-4, *C. olivaceus* 1-2, *C. planiceps*, *C. pseudoleiacanthus*, *C. pseudonieuhofii*	5.42 [3.25, 13.00]	6.37 [3.986, 9.197]	Normal (5% quantile 3.77, median 5.42, 95% quantile 7.06)
	Node 4 in Fig. 1.2a	*C. intermedius*, *C. kapuasensis*, *C. leiacanthus*, *C. meladerma* 1-3, *C. microstomus*, *C. nieuhofii* 1-4, *C. planiceps*, *C. pseudoleiacanthus*, *C. pseudonieuhofii*	3.625 [2.175, 8.70]	3.705 [2.699, 6.569]	Normal (5% quantile 1.98, median 3.63, 95% quantile 5.27)
Glyptothorax (Sisoridae, Siluriformes)	Node 1 in Fig. 1.2b	Root	3.50 [2.10, 7.70]	7.114 [4.509, 9.536]	Normal (5% quantile 1.86, median 3.50, 95% quantile 5.14)
	Node 2 in Fig. 1.2b	All MOTUs, except for *Glyptothorax botius*, *G. sp.* Buchanani-Burmanicus, *G. burmanicus*, *G. cavia*, *G. garhwali*, *G. obliquimaculatus*, *G. telchita*	3.21 [1.925, 7.70]	4.196 [2.655, 5.807]	Normal (5% quantile 1.56, median 3.21, 95% quantile 4.85)

Lineage	Calibrated node	Members	Age estimate - 1.2% clock rate (Ma)	Age estimate - StarBEAST2 (Ma)	Age estimate distribution - StarBEAST2
Dermogenys, Hemirhamphodon, Nomorhamphus (Zenarchopteridae, Beloniformes)	Node 3 in Fig. 1.2b	*G. aff. platypogon*, *G. amnestus*, *G. exodon*, *G. fuscus* 1-4, *G. nieuwenhuisi*, *G. platypogon* 1-4, *G. platypogonides* 1, *G. robustus* 1-4, *G. stibaros*	3.08 [1.85, 7.40]	1.853 [1.074, 2.628]	Normal (5% quantile 1.44, median 3.08, 95% quantile 4.73)
	Node 1b in Fig. 1.2c	*Hemirhamphodon*, except *H. tengah* & *H. chrysopunctatus*	4.29 [2.575, 10.30]	7.243 [4.578, 10.151]	Normal (5% quantile 2.65, median 4.29, 95% quantile 5.94)
	Node 2 in Fig. 1.2c	*Dermogenys* and *Nomorhamphus*	7.79 [4.675, 18.70]	6.282 [4.276, 8.656]	Normal (5% quantile 6.15, median 7.79, 95% quantile 9.44)
	Node 3 in Fig. 1.2c	*Nomorhamphus*	4.04 [2.425, 9.70]	3.903 [2.135, 5.936]	Normal (5% quantile 2.40, median 4.04, 95% quantile 5.69)
	Node 4 in Fig. 1.2c	*Dermogenys*	5.83 [3.50, 14.00]	5.886 [4.023, 8.090]	Normal (5% quantile 4.19, median 5.83, 95% quantile 7.48)
	Node 5 in Fig. 1.2c	*Dermogenys*, except *D. bruneiensis*, *D. bispina*, *D. burmanica* 1 and *D. burmanica* 2	3.00 [1.80, 7.20]	2.933 [1.985, 4.009]	Normal (5% quantile 1.36, median 3.00, 95% quantile 4.64)
Channa (Channidae, Perciformes)	Node 1 in Fig. 1.2d	*Channa bankanensis* 1-3, *C. diplogramma* 1-3, *C. lucius* 1-12, *C. micropeltes* 1-4, *C. pleurophthalma*	10.21 [6.125, 24.50]	9.030 [6.543, 11.788]	Normal (5% quantile 8.56, median 10.2, 95% quantile 11.9)
	Node 2 in Fig. 1.2d	*C. diplogramma* 1-3, *C. micropeltes* 1-4, *C. pleurophthalma*	4.625 [2.775, 11.10]	4.813 [2.973, 6.632]	Normal (5% quantile 2.98, median 4.63, 95% quantile 6.27)
	Node 3 in Fig. 1.2d	*C. baramensis*, *C. melasoma*, *C. striata* 1-13	5.54 [3.325, 13.30]	4.04 [2.826, 5.380]	Normal (5% quantile 3.90, median 5.54, 95% quantile 7.19)
	Node 4 in Fig. 1.2d	*C. andrao*, *C. aurantimaculata*, *C. bleheri*, *C. burmanica*, *C. gachua* 6, *C. pardalis*, *C. cf. stewartii*, *C. stewartii* 1-3, *C. sp.* Mogaung	7.46 [4.475, 17.90]	4.933 [3.680, 6.460]	Normal (5% quantile 5.81, median 7.46, 95% quantile 9.10)

Phylogenetic Reconstruction and Dating

Bayesian phylogenetic reconstructions were implemented with StarBEAST2 package (Ogilvie et al. 2017) of the BEAST2 suite (Heled and Drummond 2010; Bouckaert et al. 2014). This approach implements a mixed-model including a coalescent component within species and a diversification component between species that allows accounting for variations of substitution rates within and between species (Ho and Larson 2006). StarBEAST2 requires the designation of species, which were determined using the consensus of our species delimitation analyses. Best substitution models for each marker were estimated using jModelTest2 0.1.10 (Guindon and Gascuel 2003; Darriba et al. 2012). Preliminary analyses conducted on Zenarchopteridae indicated that estimating substitution model parameters jointly with node ages and topologies led to non-converging MCMC with widely fluctuating ages of the Most Recent Common Ancestor (MRCA). Hence, ML parameter estimates were used for MCMC searches in StarBEAST2 with no further estimation of the substitution

model parameters. This analytical choice constrained the number of substitution models available in StarBEAST2 and if the selected model was not supported in StarBEAST2, the next most likely model from jModelTest2 available in StarBEAST2 was used. An uncorrelated lognormal clock model was applied alongside estimation of clock rates for all analyses. The reference nodes calibrated for the species delimitation analyses were also used in the Bayesian analyses (Table 1.1) and further used with a calibrated Yule model. Age calibration was added under normal distribution priors, using a sigma of 1 in order to use wide distributions (Table 1.1). Yule birthrate value for the analysis was generated using the R-package *ape* by applying a Yule function on each ultrametric tree generated using the chronopl function for the GMYC delimitation. To yield enough statistical power (ESS values reaching >200), multiple MCMC runs with 10 million pre-burnin generations each were generated in parallel with tree states stored every 5,000 generations. Considering the number of individuals sampled and molecular heterogeneity within each group, the length of the MCMC was different for each group, ranging from 100 million for *Glyptothorax* to 250 million for *Channa*. The MCMC were run at least twice to check for the convergence of the chains and reach ESS values > 200 for MRCA age and parameters of the diversification and clock models. The statistics were visualized using Tracer 1.7.1 (Rambaut et al. 2018) to check the outputs and to determine the best converging combination of burnin for each StarBEAST2 run. Independent runs were then combined to a single concatenated dataset using LogCombiner 1.10.4 (Bouckaert et al. 2014) to compute the final Bayesian statistics and sampled trees, with or without resampling (depending on the chain lengths). The maximum clade credibility tree, age estimates and corresponding 95% highest posterior density (HPD) were summarized using TreeAnnotator 1.10.4 (Bouckaert et al. 2014).

Molecular dating constitutes a crucial step in macroevolutionary inferences (Condamine et al. 2015a) and divergence time estimates can widely vary according to calibration methods and clock models (Ho and Larson 2006; Duchêne et al. 2014). We estimated the impact of clock rates on MOTUs age estimates and diversification trends. For each clade, we rescaled the StarBEAST2 maximum credibility tree according to an assortment of molecular clocks ranging from 0.4% to 1.6% per Ma, with an increment of 0.2%, to cover a wide range of known mitochondrial substitution rates (Ortí and Meyer 1997; Hardman and Lundberg 2006; Read et al. 2006; Kadarusman et al. 2012). StarBEAST2 trees were rescaled using the Alter/Transform

Branch Lengths function in Mesquite 3.61 (Maddison and Maddison 2019) resulting in a total of seven trees computed for 0.4, 0.6, 0.8, 1.0, 1.2, 1.4 and 1.6% per Myrs clocks for each group.

Diversification Rates Estimations

We first plotted lineages through time (LTT) for each taxon using the R-package *ape* (Paradis and Schliep 2019) with confidence intervals computed with 1,000 random trees sampled from the StarBEAST2 MCMC file (Heled and Drummond 2010; Bouckaert et al. 2014). Then, to test the hypothesis that speciation was linked to past environmental changes (Condamine et al. 2013a, 2019), we designed a ML framework including five types of diversification models: constant-rate, time-dependent, temperature-dependent, sea-level-dependent, and diversity-dependent models. These models rely on the ML framework originally developed by Morlon et al. (2011) and implemented in the R-package *RPANDA* 1.3 (Morlon et al. 2016). We accounted for missing species in the phylogeny by using a global sampling fraction that is the ratio of sampled species diversity over the total described species diversity for each clade. We designed and fit a total of 17 diversification models (Table S1.1). We first fit two constant-rate models, considered as references for model comparison: one with the speciation rate constant through time with no extinction (BCST) and another with speciation and extinction rates constant through time (BCSTDCST). Second, we fit four time-dependent models: a model with only speciation rate varying through time and no extinction (BtimeVar), a model with speciation rate varying through time and constant extinction (BtimeVarDCST), a model with constant speciation and extinction rate varying through time (BCSTDtimeVar), and a model with speciation rate and extinction rate both varying through time (BtimeVarDtimeVar). We then fit eight models with speciation and extinction rates varying according to an external environmental variable of which four models had the temperature curve, and four had the sea level curve as follows: two models with speciation varying in function of the environmental variable (BtemperatureVar and Bsea-levelVar), two models with speciation varying in function of the variable with constant extinction rate (BtemperatureVarDCST and Bsea-levelVarDCST), four models with extinction rate varying as a function of the variable and constant speciation rate (BCSTDtemperatureVar and BCSTDsea-levelVar), and two models with both speciation and extinction rates varying as a function of the

variable (BtemperatureVarDtemperatureVar and Bsea-levelVarDsea-levelVar). All diversification analyses were performed independently for each the four fish clades.

We chose an exponential dependence with time, temperature or sea level, and diversification rates because this exponential function is robust and can take a broad range of shapes depending on the strength and direction of the dependence to the fitted variable. Speciation rate (λ) and extinction rate (μ) are parameterized by the set of following parameters: When speciation and extinction rates are exponential functions of the sea level through time, we used the equations $\lambda(S_{(t)}) = \lambda_0 \times e^{\alpha S(t)}$ and $\mu(S_{(t)}) = \mu_0 \times e^{\beta S(t)}$, where λ_0 and μ_0 are respectively the speciation rate and the extinction rate expected when sea level is at 0 meters. The variables α (and β) are coefficients that measure the strength and sign of the relationship with sea level (*e.g.* $\alpha > 0$ and $\beta > 0$ indicate that speciation and extinction rates increase with sea level high stands). Similar equations can be written for an exponential relationship between temperature through time and speciation and extinction rates: $\lambda(T_{(t)}) = \lambda_0 + \alpha T_{(t)}$ or $\lambda(T_{(t)}) = \lambda_0 \times e^{\alpha T(t)}$, where *T*(*t*) is the temperature at time *t* and λ_0 is the speciation rate at the temperature of 0° C.

In addition, three diversity-dependent diversification models were fitted using a maximum likelihood function with the R-package *DDD* 3.7 (Etienne et al. 2012). In the diversity-dependent models, speciation rates or extinction rates vary as a function of the number of lineages in the clade (Etienne et al. 2012). We took this function to be linear as explained in Etienne et al. (2012). The diversity-dependent models are parameterized by λ_0 and μ_0, the speciation and extinction rates in the absence of a competing lineage, and K that is the 'carrying capacity', and represents asymptotic clade size. All speciation and extinction rates were constrained to be positive. The models were compared with the corrected Akaike Information Criterion (AICc) and Akaike weight (AICω). The model with the lowest AICc and highest AICω was considered to be the best fitting model for the phylogeny. This model selection analysis was repeated for each of the seven dated trees based on different clock rate hypotheses (0.4% to 1.6% per Ma) for the groups departing from a constant diversification model and related to sea level or temperature.

Ancestral States Estimations

To infer the influence of Sundaland palaeoriver and insular systems on the diversification of Southeast Asian freshwater fishes, we estimated the ancestral areas of origin for each clade using the species trees obtained from StarBEAST2. Ancestral area estimations were performed with the R-package *BioGeoBEARS* 1.1 (Matzke 2013) using two sets of geographic delimitations based on: (1) palaeorivers, and (2) islands. Then, geographic patterns of speciation were recorded as follows: (1) no dispersal, sister species co-occur within the same palaeoriver and the same island; (2) dispersal between islands within a palaeoriver, sister species are alternatively distributed on different islands within the same palaeoriver; (3) dispersal between palaeorivers within the same island, sister species are alternatively distributed on different palaeorivers within the same islands; and (4) dispersal between islands and between palaeorivers, sister species are alternatively distributed on different palaeorivers and different islands. Ancestral area estimations involving palaeorivers were based on the following geographic areas (Fig. 1.1): (1) the North Sunda river system, (2) the East Sunda river system, (3) the Malacca Straits river system, (4) the Siam river system, (5) the Bangka-Belitung, (6) the China-South Asia, (7) the Mekong and Irrawaddy river system, (8) the Northern Borneo and (9) the Philippines. Ancestral area estimations involving islands were based on the following areas: (1) Borneo, (2) Java-Bali, (3) Sumatra, (4) Bangka-Belitung, (5) China-South Asia, (6) Mainland Southeast Asia, (7) Philippines, and (8) Sulawesi. The occurrence of each MOTU in the areas were compiled in a presence/absence matrix for each taxon and served as input data together with the species tree. The range distribution of each MOTU from both approaches is provided in Table S1.2. Ancestral area estimations were then carried out using 6 models: Dispersal-Extinction Cladogenesis (DEC) and DEC+J (Matzke 2014); ML version of Dispersal-Vicariance analysis (DIVALIKE) and DIVALIKE+J; Bayesian biogeographical inference model (BAYAREALIKE) and BAYAREALIKE+J (van Dam and Matzke 2016). The inclusion of the parameter J has been recently criticized from a conceptual and statistical perspective (Ree and Sanmartín 2018). The concept of jumping dispersal has been developed for insular systems to account for the settlement of a new lineage established by colonization without an intermediate widespread ancestor (Clark et al. 2008; Ree and Sanmartín 2018). Considering the biogeographic scenario of Sundaland and the insularity of the

system, jumping dispersal cannot be discarded *a priori* from a conceptual perspective and several studies have previously pinpoint the importance of jumping dispersal in insular systems (de Bruyn et al. 2013; Condamine et al. 2015b; Beck et al. 2017). In fact, the Palaeoriver Hypothesis predicts that sea level low stands enabled jump dispersal among islands through temporary land bridges. Models of ancestral area estimation including the J parameters were thus considered here and the best-fit model was estimated using the AICc.

Results

Phylogenetic Reconstruction and Species Delimitation

A total of 2,211 sequences from 1,511 individuals were mined from *GenBank* and BOLD, representing 110 nominal species (Table S1.2). The alignment for the genus *Clarias* consisted of 5,322 base pairs (bp), including 3,093 bp from the mitochondrial genome (16S, COI and Cytb) and 2,229 bp from the nuclear genome (RAG1, RAG2), for 146 individuals belonging to 13 morphological species. The alignment for the genus *Channa* consisted of 2,856 bp, including 2,076 bp from the mitochondrial genome (16S, COI, Cytb) and 780 bp from the nuclear genome (RAG1), for 891 individuals belonging to 30 morphological species. The alignment for the genera *Dermogenys*, *Hemirhamphodon* and *Nomorhamphus* consisted of 4,109 bp, including 984 bp from the mitochondrial genome (COI, Control Region) and 3125 bp from the nuclear genomes (DP5, DP14, DP21, DP35, DP37, HP5, HP54, HPR56), for 266 individuals belonging to 17 morphological species. The alignment for the genus *Glyptothorax* data consisted of 2,776 bp, including 1,869 bp for the mitochondrial genome (COI, Cytb) and 907 bp for the nuclear genome (RAG2), for 208 individuals belonging to 50 morphological species. All alignments are available in TreeBASE (TB2:S26912).

Substitution models for each marker are provided in Table S1.3 (Online supplementary material). The four ML phylogenetic reconstructions generally provided well-supported clades with most internal nodes supported by BP > 80 except within the genus *Channa* (Fig. 1.2, Fig. S1.1). For the genus *Clarias*, Asian species constitute a monophyletic group (Fig. 1.2a, clades II + III) separated from the African species (Fig. 1.2a, clade I), with *Clarias gariepinus* (Burchell 1822) as sister to all remaining species. All species are monophyletic except *C. nieuhofii* Valenciennes 1840, which is

recovered as polyphyletic with three distinct lineages (Fig. S1.1). For the genus *Glyptothorax*, the inferred tree is imbalanced with continental species constituting a stem group at the root of the tree (Fig. 1.2b, Fig. S1.1) and Sundaland species constituting a monophyletic group (Fig. 1.2b, clade I). All species are monophyletic, except *G. platypogonides* (Bleeker 1855) and *G. fuscus* Fowler 1934, which are polyphyletic (Fig. S1.1). The monophyly of the Zenarchopterid genera *Hemirhamphodon* (Fig. 1.2c, clade I) *Nomorhamphus* (Fig. 1.2c, clade II), and *Dermogenys* (Fig. 1.2c, clade III) is recovered with a sister-relationship between *Dermogenys* and *Nomorhamphus* (Fig. 1.2c, clades II + III). All species are monophyletic, except *Nomorhamphus megarrhamphus* (Brembach 1982), which is paraphyletic (Fig. S1.1). For the genus *Channa*, internal relationships are poorly supported, but four main clades are observed (Fig. 1.2d, clades I to IV). All *Channa* species are monophyletic, except *Channa gachua* (Hamilton 1822), which is polyphyletic and includes at least three lineages distinctly associated to other species, and *C. lucius* (Cuvier 1831) and *C. marulius* (Hamilton 1822), which are paraphyletic (Fig. S1.1).

As expected, different species delimitation analyses yielded variable number of MOTUs (Table S1.2): (1) 47 *Clarias* using mGMYC, 32 using PTP, 7 using RESL and 24 using ABGD, resulting in a consensus of 29 MOTUs, (2) 107 *Glyptothorax* using mGMYC, 58 using PTP, 58 using RESL and 72 using ABGD, resulting in a consensus of 61 MOTUs, (3) 48 Zenarchopteridae using mGMYC, 27 using PTP, 35 using RESL and 49 using ABGD, resulting in a consensus of 43 MOTUs, and (4) 521 *Channa* using mGMYC, 120 MOTUs using PTP, 57 using RESL and 133 using ABGD resulting in a consensus of 120 MOTUs. The consensus delimitation scheme resulted in between 1.22 (*Glyptothorax*) and 4.00 (*Channa*) more MOTUs than morphological species (Table S1.2).

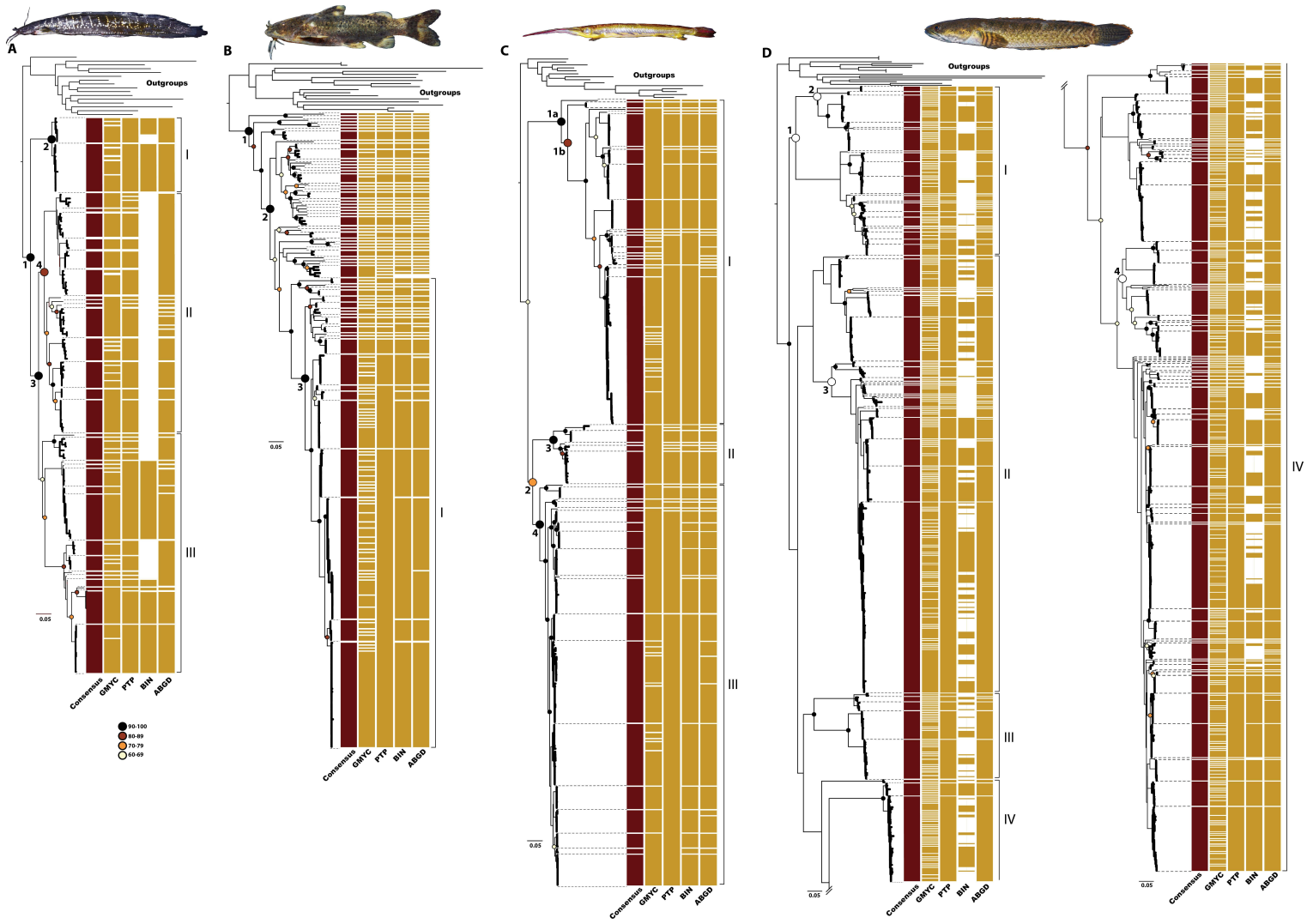

Figure 1.2 Maximum likelihood trees and species delimitation for (A) *Clarias*, (B) *Glyptothorax*, (C) Zenarchopteridae and (D-E) *Channa*
For each clade, MOTUs delimited according to GMYC, PTP, BIN and ABGD (yellow), and the 50 percent consensus scheme of delimitation (red) are shown. Node circles represent bootstrap proportions and nodes numbering corresponds with the calibration nodes in Table 1.1. Major clades are labeled on the right.

The final Bayesian analyses as implemented in StarBEAST2 were performed using MCMC of length required to reach convergence (ESS > 200). For *Clarias,* two parallel runs were run with a burnin of 65.01 million generations and 15 million generations, respectively, and both runs were concatenated into a 220-million-generation dataset of 44,000 sampled trees. For *Glyptothorax,* three parallel runs were launched with burnins of 5, 10 and 5 million generations, respectively, and then concatenated into a 280-million-generation dataset with 56,000 sampled trees. For *Channa,* four parallel runs were used with burnins of 182.5, 27.52, 47 and 43 million generations, respectively, and concatenated into a 700-million-generation dataset with 35,000 sampled trees. For Zenarchopteridae*,* four parallel runs were used with burnin of 70, 50, 38 and 60 million generations, respectively, and concatenated into a 182-million-generation dataset with 18,200 sampled trees. The StarBEAST2 chronograms were largely congruent with the topologies estimated with the ML approach (Fig. S1.2). The ages of the MRCA vary between clades with a MRCA dated at 7.09 Ma (95% HPD = 4.49-10.31 Ma,) for *Clarias*, 7.1 Ma (95% HPD = 4.51-9.54 Ma) for *Glyptothorax*, 12.63 Ma (95% HPD = 8.24-18.68 Ma) for Zenarchopteridae, and 14.35 Ma (95% HPD = 10.54-18.41 Ma) for *Channa* (Fig S1.2). For *Clarias*, the MRCA of the Asian species is dated around 6.37 Ma (95% HPD = 3.99-9.19 Ma) and the Zenarchopteridae genera are dated around 9.88 Ma (95% HPD = 6.02-14.70 Ma) for *Hemirhamphodon*, 5.88 Ma (95% HPD = 4.02-8.09 Ma) for *Dermogenys*, and 3.90 Ma (95% HPD = 2.14-5.94 Ma) for *Nomorhamphus*.

Diversification Rates

Most of the MOTUs delineated in the consensus scheme originated in the Pleistocene with divergence time estimates younger than 2.5 Myrs (Fig. 1.3). More than half (53%) of the morphological species and 88% MOTUs are younger than 2.5 Myrs (Fig. 1.3a). The LTT plots indicate that 84% of extant fish diversity occurred during the Pleistocene. On Sundaland, Pleistocene speciation peaks during sea level low stands with 88% of speciation events inferred during the last 2.5 million years and 76% during the last 1.5 million years. This peak of Pleistocene speciation is observed for a wide range of clock rates, including some of the slowest clock rate hypothesis with 72% and 57% of speciation events during the last 2.5 and 1.5 Myrs, respectively, for the 0.6% per millions years (Fig. S1.3). Faster rates yielded a higher percentage of

young MOTUs. The 1.6% per Myrs rate yielded 93% and 83% of MOTUs younger than 2.5 and 1.5 Myrs, respectively. The only exception was observed for the 0.4% per million years with 60% and 38% of the MOTUs younger than 2.5 and 1.5 Myrs, respectively.

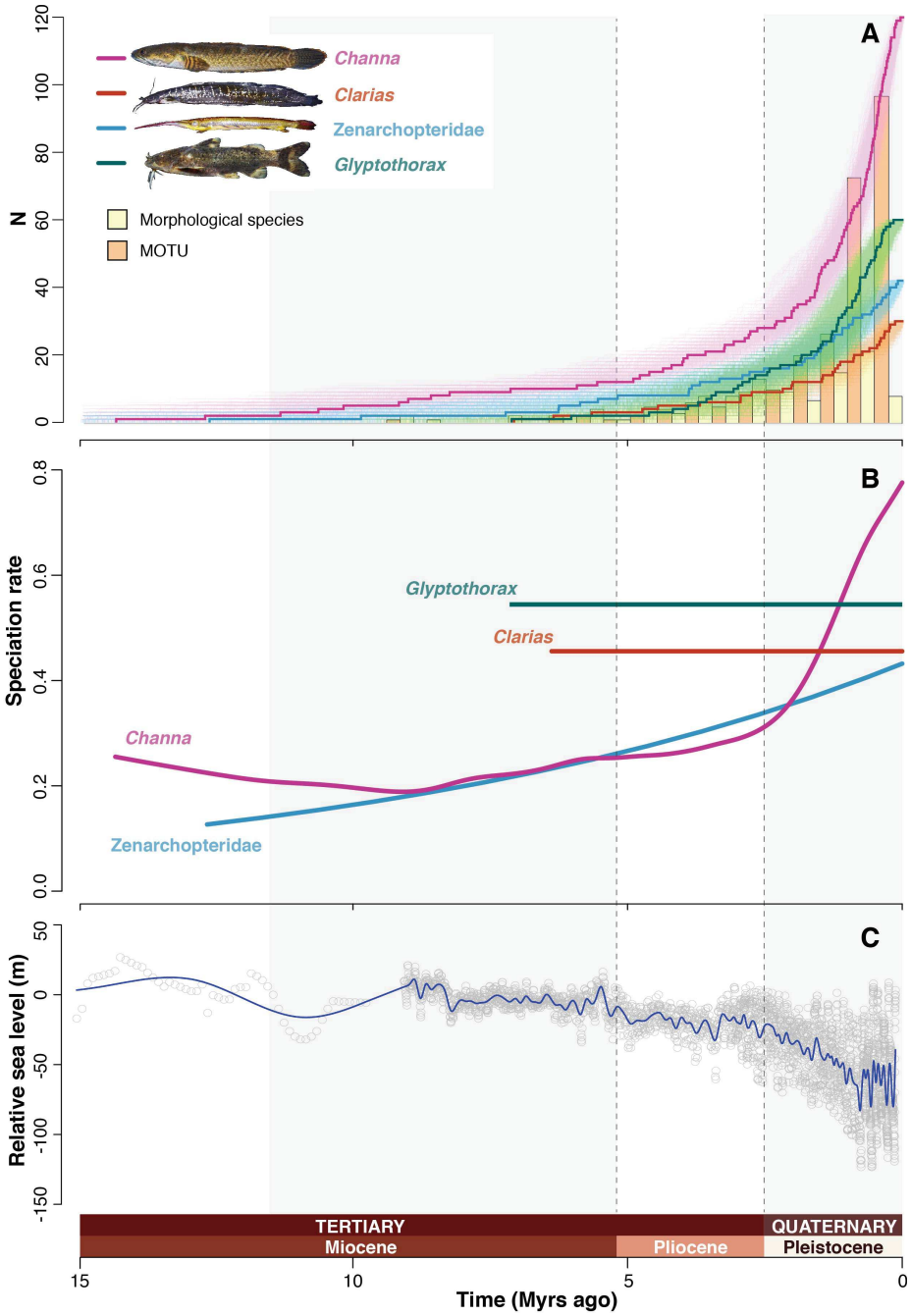

Figure 1.3 Diversification through time of Sundaland freshwater fishes. Panel A shows the lineages through time (LTT) plot with confidence intervals collected from 1000 randomly sampled trees along the StarBEAST2 MCMC, superimposed with the total number of nominal species and MOTUs through time (0.5 Myrs class). Panel B shows the speciation rate through time for each group, based on the most likely model of diversification (constant for Clarias and Glyptothorax, time-exponential for Zenarchopteridae, sea level-dependent for Channa). Panel C shows the sea level fluctuations over the last 15 Myrs (adapted from Miller et al. 2005)

Likelihood scores and AICω for 17 diversification models indicate clade-specific patterns of diversification (Table S1.4). For *Clarias* and *Glyptothorax*, the constant-rate speciation model is the most likely according to AICω, respectively 0.267 and 0.212, with high speciation rates (λ) of 0.4556 and 0.5444, respectively. For Zenarchopteridae, three models are equally likely, including: (1) the time-dependent speciation model (AICω = 0.145, λ_0 = 0.432) with a speciation increasing through time (α = -0.0971); (2) the constant-rate speciation model (AICω = 0.119, λ_0 = 0.3316); and (3) the temperature-dependent speciation model (AICω = 0.116, λ_0 = 0.5132) with speciation increasing as temperatures cooled (α = -0.1444). While diversification patterns of the previous three groups show no dependency on sea-level eustasy, the most likely model for *Channa* diversification through time is the sea-level-dependent speciation model (AICω = 0.415, λ_0 = 0.2279) with speciation increasing as sea level dropped (α = -0.0182). These clade-dependent diversification patterns are reflected by the heterogeneous trends of speciation rates through time for the four lineages (Fig. 1.3b), including constant speciation rates through time for *Glyptothorax* and *Clarias*, a sea-level dependent speciation rate for *Channa*, and a time-dependent speciation rate for Zenarchopteridae. The dependency of *Channa* diversification on sea level eustasy was recovered for all clock rates except for the two slowest clocks (Table S1.5). The most likely model for the 0.4% per Myrs rate includes a positive relationship between speciation and sea level, speciation increased as sea level increased (α = 0.0119). The most likely model for the 0.6% per Myrs clock includes a positive relationship between extinction rates and sea levels eustasy, extinction increased when sea level increased (β = 0.0351).

Geography of Diversification

Ancestral area estimations for all groups with either island-based or palaeoriver-based result in higher likelihood for models incorporating the J parameter (Table S1.6). The DEC+J model is the most likely, and was further used in subsequent analyses. Most of the colonization generally initiated before the Pleistocene and in Borneo (the North Sunda and East Sunda River Systems, Fig. 1.4), while several subsequent colonizations happened through Sumatra, and Java is typically the last island colonized. This pattern was observed within *Clarias*, with a first vicariance event inferred between Borneo and the Southeast Asian mainland at 6.37 Ma (95% HPD =

3.99-9.19 Ma) and associated with the first split in the Asian *Clarias* (Fig. 1.4a, clade II vs clade III), the colonization of Sumatra happening subsequently twice from Borneo (Fig. 1.4a, clade III). The same pattern is observed for *Glyptothorax* with a vicariance inferred between Borneo and mainland around 3.64 Ma (95% HPD = 2.34-5.17 Ma; Fig. 1.4b, clade I). Different patterns are observed for Zenarchopteridae (Fig. 1.4c) and *Channa* (Fig. 1.4d), indicating that the islands of Sundaland were colonized multiple times.

Most speciation events are inferred to happen within islands, as exemplified by the diversification of three of the four lineages (*Clarias*, *Glyptothorax* and Zenarchopteridae) in Borneo and Sumatra (Figs. 1.4a, 1.4b and 1.4c) or Java for *Channa* (Fig. 1.4d). The same trend is observed for palaeorivers; most speciation events are inferred to occur within them (Fig. 1.4). Four major geographic patterns of speciation were identified: (1) sister species occupy the same island and the same palaeoriver, *e.g*. *Glyptothorax platypogon* 1-4 from Java (East Sunda River System), (2) sister species occupy different islands but the same palaeoriver, *e.g*. *Glyptothorax robustus* (Boeseman 1966) 1-4 from Java and Sumatra (East Sunda River System), (3) sister species occupy different palaeorivers within the same island, as in *G. amnestus* (North Sunda) and *G. fuscus* 4 (Malacca Strait), and (4) sister species occupy different islands and different palaeorivers *e.g*. *Clarias pseudoleiacanthus* Sudarto, Teugels and Pouyaud 2003 (North Sunda in Borneo) and *C. leiacanthus* Bleeker 1851 (Malacca Strait in Sumatra).

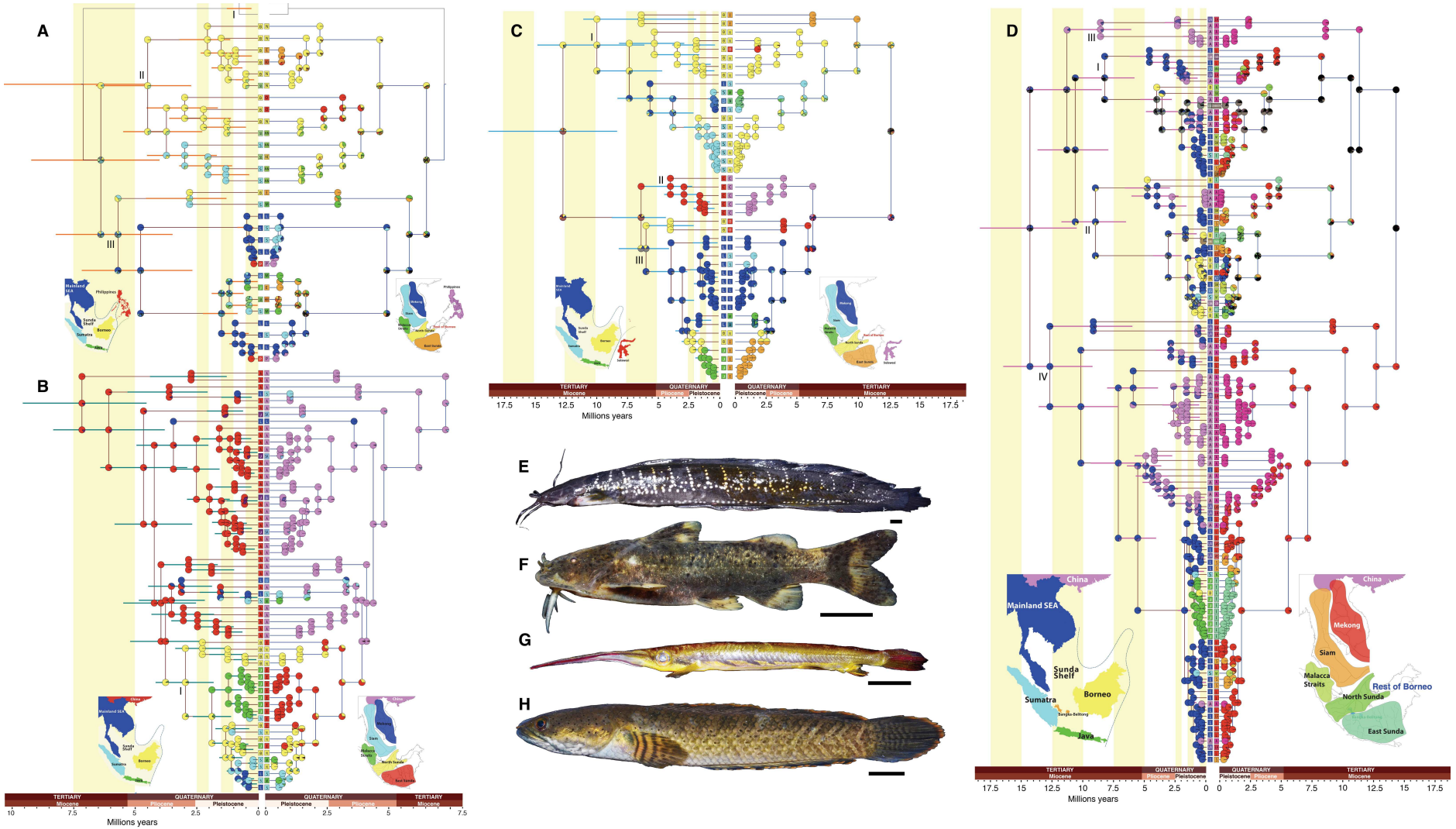

Figure 1.4 Panels A to D show the Bayesian maximum credibility trees for (A) *Clarias*, (B) *Glyptothorax*, (C) Zenarchopteridae and (D) *Channa* as well as the ancestral area reconstructions for each group, based on islands (left) and palaeorivers (right). Panels E to F show exemplary specimens for (E) *Clarias*, (F) *Glyptothorax*, (G) Zenarchopteridae and (H) *Channa* with their relative size (1 cm scale).

In total, 66.3% of the speciation events inferred involve dispersal either between islands or between palaeorivers (Table 1.2). In Sundaland, 59.2% of speciation events occurred within island (40.8% between islands) while 39.8% of speciation events occurred within palaeorivers (60.2% between palaeorivers). These trends are observed for all groups except Zenarchopteridae, which speciation events within islands represent 81.7% and speciation within palaeorivers 63.6% (Table 1.2). The proportion of speciation events within or between islands is stable through time for speciation events involving dispersal between palaeorivers (Fig. 1.5a). However, this pattern was not observed for Zenarchopteridae (Fig. 1.5b) with a transition of dispersal between palaeorivers within to between islands. In *Glyptothorax*, most speciation events between palaeorivers involve dispersal between islands (Fig. 1.5e). Within palaeorivers, speciation events involving dispersal between islands are scarce for all groups (Figs. 1.5b, c, d, e).

Table 1.2 Summary statistics of geographic patterns of speciation events for *Clarias*, *Glyptothorax*, Zenarchopteridae and *Channa*

Taxa	Islands		
	Within (%)	Between (%)	Total (%)
All	59.2	40.8	-
within palaeorivers	33.7	6.1	39.8
between palaeorivers	25.5	34.7	60.3
Clarias	55.6	44.4	-
within palaeorivers	22.3	11.1	33.4
between palaeorivers	33.3	33.3	66.6
Glyptothorax	50	50	-
within palaeorivers	40.9	4.5	45.4
between palaeorivers	9.1	45.5	54.6
Zenarchopteridae	81.7	18.3	-
within palaeorivers	54.5	9.1	63.6
between palaeorivers	27.2	9.1	36.4
Channa	52.8	47.2	-
within palaeorivers	22.2	2.8	25
between palaeorivers	30.6	44.4	75

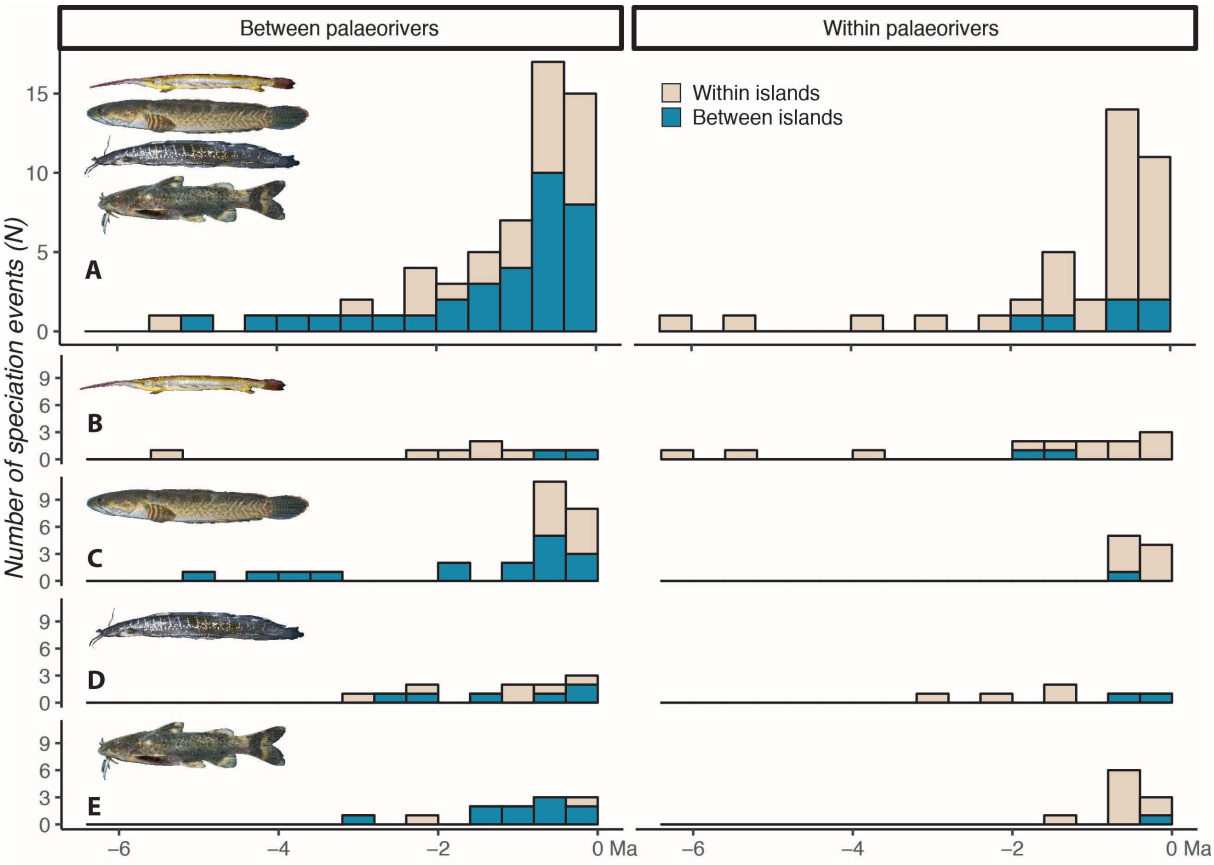

Figure 1.5 Geographic pattern of speciation in Southeast Asian freshwater fishes. The plots show the numbers of speciation events associated to the four geographic patterns of speciation as follows: between islands and within palaeorivers, between islands and between palaeorivers, within islands and within palaeorivers, and within islands and between palaeorivers. Each bar shows the number of speciation events for each category for (A) all groups (cumulative), (B) Zenarchopteridae, (C) *Channa*, (D) *Clarias* and (E) *Glyptothorax*.

Discussion

Diversification of Southeast Asian Aquatic Biotas

The "Late Pleistocene" (Barraclough and Nee 2001; Wiens and Donoghue 2004; Mittelbach et al. 2007; Beck et al. 2017) and the "Pleistocene Species Pump" (Esselstyn and Brown 2009; Brown et al. 2013; Papadopoulou and Knowles 2015a, 2015b; Li and Li 2018) hypotheses suggest that sea-level fluctuations accelerated diversification of insular biotas. However, we found no signature of sea-level fluctuations on diversification trends except for one of the four groups. The genus *Channa* is the only group in which diversification rate was associated to sea-level fluctuations: its speciation rate increased as sea level dropped. This translates into increasing speciation rates through time since sea level generally decreased through the Pleistocene (Miller et al. 2005). Hence, the genus *Channa* had the highest

speciation rate of all studied taxa at present (Fig. 1.3b). The superior dispersal ability and environmental adaptability of *Channa* compared to the other taxa may enable them to colonize many different types of environments. With the closure of dispersal routes between islands during sea level highstands, gene flow is reduced, thereby increasing the probability of divergence between populations and eventually speciation. Such speciation events likely occurred repeatedly over short time intervals (glaciation periods). Hence, the genus *Channa* is the only taxon we studied that supports the "Pleistocene Species Pump" hypothesis.

On the contrary, the diversification of the other clades does not support this hypothesis. Both silurid groups (*Clarias* and *Glyptothorax*) diversified at a constant rate despite different life history strategies. *Glyptothorax* had a higher speciation rate than *Clarias*, which might be related to the tendency of *Glyptothorax* to occupy hilly streams that are more isolated (Hutama et al. 2017). The dispersal abilities of *Clarias* likely limited allopatric speciation. Such specificity for *Glyptothorax* lineages could have (1) inhibited inter-island dispersal during low sea level since the freshwater corridors between islands were unsuitable, as well as (2) facilitated isolation and *in situ* diversification due to restricted gene flow (Hubert et al. 2015a; Hutama et al. 2017), hence generating high, constant speciation rate. In the same way, none of the diversification trends for Zenarchopteridae were directly related to sea-level fluctuations.

Diversification trends show no tendency to decline following a diversity-dependent diversification (DDD) model (Seehausen 2007; Rabosky and Lovette 2008; Seehausen et al. 2008). The DDD model is based on the assumption that biotas are bounded by the carrying capacities of the environment, either in terms of habitat diversity (Kisel et al., 2011; Phillimore & Price, 2008; Schluter, 2000) or the total number of individuals that can be sustained (Alonso et al., 2006; Hubbell, 2001). This model predicts that diversification rates slowdown through time as a consequence of (1) declining speciation rates due to the increasing occupation of available niches by proliferating species (Schluter 2000; Gavrilets and Vose 2006; Hubert et al. 2015a), or (2) increasing extinction rate through time over constant speciation rate as a consequence of the increasing importance of stochastic demographic dynamics within proliferating species with ever reduced population size (Alonso et al. 2006; Rabosky and Lovette 2008; Hubert et al. 2015a). Here, however, we found no trend for declining diversification rates in any taxa. Two of the four clades have rates that increase through

time (Fig. 1.3b). The high diversity of freshwater habitats in Sundaland likely provides a diversity of niches to sustain elevated and/or constant diversification. Water chemistry is widely diverse in tropical systems (Guyot 1993) and Sundaland is no exception with water types ranging from clear and alkaline waters in karts (Clements et al. 2006), turbid, sediments rich water of the floodplains to the acidic, humic acid rich waters of the peat swamps (Ng et al. 1994; Giam et al. 2012). The high number of species distributed among multiple palaeorivers with sister species diversifying within the same palaeoriver seems consistent with the existence of such high diversity of habitat/niche within palaeorivers. However, cases of ecological speciation across freshwater habitats are scare for riverine habitats (Sullivan et al. 2002; Nolte et al. 2005; Alter et al. 2017), most cases have been discovered in ancient lakes or closed systems (Nagel and Schluter 1998; Verheyen et al. 2003; Barluenga and Meyer 2004; Herder et al. 2006; Landry et al. 2007). The Zenarchopteridae and *Clarias* species are distributed across a wide range of water types, with some *Clarias* and *Hemirhamphodon* species occupying acidic waters of peat swamps while other *Clarias* and *Dermogenys* species occupy clear waters and floodplains (Pouyaud et al. 2009; Lim et al. 2016a; Nurul Farhana et al. 2018). However, water type transitions are scarce in their phylogenies with peat swamps and floodplain-associated species of *Clarias* grouped in distinct clades. This trend suggests that adaptive divergence across freshwater habitats is limited in the studied groups. By contrast, landscape fragmentation by geology has been widely document in the area (Nguyen et al. 2008; Lim et al. 2016a; Dahruddin et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018; Hubert et al. 2019).

The habitat diversity within palaeorivers might be expected to have interacted with the availability of dispersal routes among palaeorivers, which was highly affected by sea-level fluctuations (Brown et al. 2013). Similar trends have also been reported for the Pleistocene Aggregate Island Complex (PAIC) model in the Philippines archipelago. Brown et al (Brown et al. 2013) which suggested an alternative "nested PAIC model" considering the interplay of the external sea level fluctuations causing cycles of island connection-isolation and changing local habitat diversity within an island. Due to both its palaeoriver and insular systems, Sundaland is likely a comparable system with large and highly heterogeneous freshwater habitats (Brown et al., 2013; Esselstyn & Brown, 2009; Papadopoulou & Knowles, 2015b, 2015a). The frequent regional intermediate disturbance associated to PCF likely impacted biotic

interactions among Southeast Asian freshwater fish species and probably triggered a reshuffling of their ecological communities, which may explain the higher species richness than expected under a DDD model.

Diversification Mostly Occurred Within Islands and Palaeorivers

Our ancestral area estimations show that diversification frequently resulted from long-distance dispersal followed by *in situ* diversification (founder effect speciation) as suggested by the high likelihood of the models with the J parameter (Matzke 2014). This pattern has been previously proposed through detection of trans-island dispersal of freshwater fishes (Dodson et al. 1995; Pouyaud et al. 2009; Adamson et al. 2010; Tan et al. 2012; Tan and Lim 2013; de Bruyn et al. 2013; Hubert et al. 2015a; Lim et al. 2016a; Beck et al. 2017; Dahruddin et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018; Hubert et al. 2019). The establishment of taxa from mainland Asia during the Miocene and Pliocene supports the pre-Pleistocene colonization hypothesis (Dodson et al. 1995; de Bruyn et al. 2013; Hendriks et al. 2019). These results agree that regional shallow seas only started to dry out during the Pliocene, offering the possibility for dispersal of freshwater ancestors from mainland Asia. During these periods, Borneo was certainly more connected to mainland Southeast Asia than Sumatra and Java (Figs. 1.1b & 1.1c), and Borneo probably played a role in the initial diversity build-up of Sundaland biotas (de Bruyn et al. 2013, 2014). From our inferences, Sumatra and Java were first colonized from Borneo around 2.96 Ma via the North Sunda river system and around 1.18 Ma via the East Sunda river system, respectively. Thus, our study further supports Borneo as the most origin of insular Sundaland freshwater fish diversity as previously suggested (de Bruyn et al. 2014).

Although the fish lineages we studied were present in Sundaland before the Pleistocene, most of the MOTUs delimited here result from Pleistocene diversification events that are expected to be affected by PCF (Esselstyn et al. 2009). Looking at the geography of speciation for each group, we estimated that speciation of *Clarias* and *Channa* involved more dispersal events compared to *Glyptothorax* and Zenarchopteridae, probably due to higher dispersal ability (Kottelat et al. 1993; Berra 2001; Meisner 2001; Pouyaud et al. 2009; Adamson et al. 2010; Jiang et al. 2011; Tan and Lim 2013; Serrao et al. 2014; Ng and Kottelat 2016; Lim et al. 2016a; Conte-Grand et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018). The dispersal patterns and

absence of decline in the diversification rates suggest that land bridges and the subsequent Pleistocene eustatic fluctuations drove further *in situ* diversification with the interplay of the insular and palaeoriver watersheds boundaries, as well as habitat heterogeneity, which likely conditioned dispersal. Episodes of sea level rise during interglacial periods might be expected to have fragmented Sundaland. Rising seas created refugial insular areas in Borneo, Sumatra and Java. These saltwater barriers may have driven further within-island diversification explaining the numerous local radiations/*in situ* diversification events after long-distance dispersals in each of the four lineages. Yet, nearly half of Sundaland Pleistocene speciation events involve dispersal between different islands. One possible explanation is to consider the restored interconnectivity of both Sundaland land bridges and palaeoriver systems during Pleistocene glacial periods. Considering cooler and drier climate during Pleistocene glaciations, it has been proposed that savanna and seasonal forest corridors expanded through the interior of Sundaland, notably in East Sunda (Heaney 1991; Bird et al. 2005). Such inter-island bridges might be not easily penetrable by evergreen forest dependent taxa (Heaney 1991; Gorog et al. 2004; Bird et al. 2005; Pouyaud et al. 2009; Wurster et al. 2019). For example, despite its exemplary ability to disperse, *Clarias* was more likely to speciate within an island (55.6%) than between islands (44.4%), which is supported by the general phylogenetic division of the genus between black-water lineages and white-clear water lineages. Black water lineages of *Clarias* are unlikely to switch habitats, inhibiting them to easily penetrate freshwater habitats within the savanna/seasonal forest corridors during glaciation periods (Pouyaud et al. 2009). Zenarchopteridae was the lineage least likely to disperse between islands (18.3 %), most probably due to its limited larval dispersal ability of these mostly ovoviviparous fishes (de Bruyn et al. 2013). Inter-island dispersal seems to be correlated not only with dispersal ability, but also with habitat specificity (Heaney 2007).

The substantial proportion of speciation with dispersal between islands coupled with dispersal between palaeoriver systems and of speciation between palaeorivers argues in favor of the importance of habitat heterogeneity/diversity in the diversification of Sundaland freshwater fishes. If palaeoriver systems act as dispersal channels, within or between islands, one can also argue that borders of both palaeorivers and insular systems might not be as clearly delineated as we thought, due to biological aspects including the difference in vegetation cover, and/or physical aspects such as island geomorphology (Pouyaud et al. 2009; Brown et al. 2013; Hutama et al. 2017).

The Bangka-Belitong islands are located at the boundary between North Sunda and East Sunda river systems. We found no clear signal of its affinity for one of the two palaeorivers. Similarly, borders between the two river systems also could not be recovered strictly by biogeographic estimates in Lampung (eastern Sumatra) and the southwestern part of Borneo, probably because the lowland freshwater habitat topography facilitates gene flow between them. Notable examples for the unclear boundaries between North Sunda and East Sunda river systems are exemplified by: (1) the existence of sister lineages of *G. platypogonides* 1 and *G. stibaros* (Borneo), (2) dispersal during cladogenesis between the *G. pictus* and *G. major* groups in Borneo, and (3) the trans-palaeoriver distributions of *Clarias leiacanthus* (Sumatra) and *C. meladerma* 1 (Sumatra and Borneo). The physical island geomorphology could also contribute to re-arrangements of palaeoriver watersheds limits through headwater capture events among palaeorivers through time. The Bornean part of the North Sunda and East Sunda river systems share the same headwater area in the center of the island.

Habitat fragmentation by mountainous terrain has been identified as a major geomorphological driver for both *in situ* diversification and abundant cryptic diversity in Java (Dahruddin et al. 2017; Hutama et al. 2017; Hubert et al. 2019). The existence of a single palaeoriver system may have driven ancestral lineages on the island to diversify further within isolated pockets of riparian environment. This suggests that extant freshwater fish diversity in Java resulted from high level of *in situ* diversification rather than immigration from elsewhere (Nguyen et al. 2008; Pouyaud et al. 2009; Hubert et al. 2015a; Kusuma et al. 2016; Dahruddin et al. 2017; Hutama et al. 2017; Hubert et al. 2019). Such *in situ* diversification within islands are found for all lineages studied here (e.g. *Glyptothorax platypogon*, *G. robustus*, *Dermogenys pusilla*, and *Channa gachua*; Fig. 1.4, Fig. S1.2, Table S1.2).

Robustness of the Inferences, Limits and Perspectives

For all four taxa under study, we generally found that tree topologies are congruent with previously published phylogenetic hypotheses (Pouyaud et al. 2009; Jiang et al. 2011; de Bruyn et al. 2013; Conte-Grand et al. 2017). Both ML and Bayesian analyses retrieved similar and robust topologies, and each genus was found to be monophyletic, including the three genera of Zenarchopteridae, with

Nomorhamphus being the sister group of *Dermogenys* and *Nomorhamphus* + *Dermogenys* as the sister group of *Hemirhamphodon,* supporting Meisner (2001) and Lovejoy et al. (2004). Minor differences concern the monophyly of *Dermogenys*, well supported now, contrary to previous studies (de Bruyn et al. 2013), and likely resulting from the incorporation of multiple outgroups here. For *Clarias*, our analysis supports the monophyly of Asian *Clarias* as previously suggested (Pouyaud et al. 2009). By contrast, ML phylogenetic reconstruction (Fig. 1.2a, Fig. S1.1) and the Bayesian species tree (Fig. 1.4a, Fig. S1.2) support the reciprocal monophyly of black water species (Clade II) and white and clear water species (Clade III). Only *Clarias microstomus* Ng 2001 and *C. planiceps* Ng 1999 departed from this general trend. For *Glyptothorax*, differences with the reference phylogeny (Jiang et al. 2011) are only due to a recent revision of Sundaland *Glyptothorax* species that led to the revision and description of new taxa (Ng and Kottelat 2016), that have been incorporated here as well as the addition of supplementary sequences (Hutama et al. 2017). Our findings are consistent with the occurrence of two distinct clades in Sundaland (Fig. 1.2b, Fig. 1.4b, Fig. S1.1-S1.2) as suggested earlier (Jiang et al. 2011). Finally, the species tree of *Channa* recovered here is consistent with the eight species groups previously recognized (Rüber et al. 2020). Our results differ from previous studies in placing the *C. punctata* group and *C. gachua* group in early-diverging position instead of the *C. micropeltes* group and *C. lucius* group (Conte-Grand et al. 2017; Rüber et al. 2020).

At the MOTUs level, several nominal species are not monophyletic, which is probably due to practical taxonomic limitations in which different lineages with no apparent morphological differences might have been lumped together into the same nominal species as in the polyphyletic *Clarias nieuhofii* (as previously observed by Pouyaud et al. 2009), *Glyptothorax platypogonides* and *Channa gachua*, as well as the paraphyletic *Nomorhamphus megarrhamphus* or several *Glyptothorax* species (Jiang et al. 2011; Ng and Kottelat 2016). These results are consistent with recent findings about the existence of high levels of cryptic diversity among Sundaland freshwater fishes (Nguyen et al. 2008; Pouyaud et al. 2009; Hubert et al. 2015a; Conte-Grand et al. 2017; Dahruddin et al. 2017; Hutama et al. 2017; Hubert et al. 2019; Sholihah et al. 2020) which further points to the necessity of using genetic delimitation methods for subsequent macroevolutionary analyses of complex biotas (Esselstyn et al. 2009; Patel et al. 2011; Ruane et al. 2014; Hubert et al. 2015a; Hutama et al. 2017; Hubert et al. 2019). The remarkable number of MOTUs recovered from nominal

species (252 MOTUs vs. 110 morphological species) calls for more detailed taxonomic works on these taxa (Pouyaud et al. 2009; Hutama et al. 2017).

The divergence times we estimated using a molecular clock approach agree with (*Hemirhamphodon, Dermogenys, Nomorhamphus*) or are younger than (*Clarias* and *Channa*) previous estimates based on a geological and fossil calibrations. We also provided the first time-calibrated phylogeny for *Glyptothorax*. We found similar divergence times for *Hemirhamphodon* (9.88 Ma, 95% HPD = 6.02-14.70 Ma, vs. 11.2 Ma, 95% HPD = 6.7-16 Ma; de Bruyn et al. 2013) and for the *Dermogenys* + *Nomorhamphus* clade (6.28 Ma, 95% HPD = 4.28-8.66 Ma, vs. 13.3 Ma, 95% HPD = 8-18 Ma; de Bruyn et al. 2013). Younger divergence estimates are found in *Clarias* and *Channa*: we estimated the Asian *Clarias* clade to be 6.37 Ma (95% HPD = 3.99-9.19 Ma) while Pouyaud et al. (2009) dated the divergence at 33.4 Ma. In the same way, we found a younger divergence time of for *Channa* (14.35 Ma, 95% HPD = 10.54-18.41 Ma) than a previous estimate (28 Ma, 95% HPD = 24-32 Ma; Rüber et al. 2020). A potential cause of these discrepancies is likely the use of a mixed coalescent/diversification model here for the Bayesian species trees reconstruction that accounts for the heterogeneity of absolute substitution rates within and between species (Ho and Larson 2006). Within species, clock rates can be considerably higher than between species because observed substitution rates are close to mutation rates, genetic drift being later responsible for the loss of haplotypes leading to stationary substitution rates (Ho and Larson 2006) that are most commonly used in phylogenetic reconstructions and molecular age estimates. Along the same line, the use of alternative clock rates moderately impacted our main results: (1) most MOTUs originated during the Pleistocene whatever the clock rate used; (2) diversification model selection was consistent across a wide range of clock rates. Furthermore, this study is based on genetic recognition of MOTUs instead of morphological species; the high proportion of cryptic diversity found here was not properly accounted for by the previous phylogenetic reconstructions (Pouyaud et al. 2009; de Bruyn et al. 2013). It can be expected that such a difference in taxon sampling has impacted absolute age estimates. The general concordance in the distribution through time of MOTUs age among taxa with different life history traits and origins argue in favor of the robustness of the age estimates established here.

Finally, several issues can be highlighted that warrant further studies. First, although the samples generally cover a broad range of localities around Sundaland,

we identify sampling gaps in large areas, including Sumatra for all taxa and in Peninsular Malaysia for *Clarias*. Second, no Barcode Index Numbers (BINs) have been assigned to COI sequences of older datasets (*Clarias* and *Channa*). While there was no significant challenge from missing BINs during species delimitation, more complete BIN would increase confidence in the species delimitation results. The subfamily Rasborinae (Brittan 1972; Sholihah et al. 2020) could be an additional biological model to assess the evolutionary history of Sundaland freshwater fishes. It has multiple species-rich genera of small size (Liao et al. 2011; Tan and Armbruster 2018). Lastly, the constant diversification rates through time inferred for two genera might result from inadequate sampling. It has also been suggested that the DDD model might be biased toward spurious detection of early explosive speciation dynamics resulting from the underestimation of the actual number of evolutionary lineages created by taxonomic and sampling bias (Barraclough and Nee 2001; Rabosky and Lovette 2008). Alternatively, constant diversification rates has been a recurrent finding in tropical biomes (Esselstyn et al. 2009; Patel et al. 2011). In this case, we have emphasized our effort to avoid clade-specific taxonomic hypotheses (Esselstyn et al. 2009) by utilizing species delimitation methods for all clades. We have detected numerous cryptic lineages for all taxa which might explain the constant diversification rate found here, as proposed by Patel et al. (2011) to explain the constant diversification rates of Neotropical *Pteroglossus*.

Conclusions

Our results indicate that the diversification and biogeography of freshwater fishes on Sundaland are not solely dependent on Pleistocene sea-level fluctuations and associated palaeoriver systems, but are also affected by: (1) pre-Pleistocene geological history of Sundaland, both at the origin of the development of palaeoriver systems and responsible for ancient speciation events among palaeorivers; (2) the idiosyncratic effects of island boundaries and the extent of palaeorivers on diversification in each system; (3) ecology of the palaeoenvironments (Heaney 2007) and the (re)emerged palaeorivers that were running through the land bridges of Sundaland through the Pleistocene (Heaney 1991; Bird et al. 2005; Slik et al. 2011); (4) the geomorphology of emerged and currently submerged Sundaland land masses, creating vague borders among palaeoriver systems; and (5) different evolutionary

responses of different groups with specific life history traits. Furthermore, it has been suggested recently that the subsidence of Sundaland may have triggered the merge of Sundaland landmasses during glacial times only 400,000 years ago, implying that eustatic fluctuations prior to that period only marginally impacted the extent of emerged lands (Husson et al. 2019). Although, not in line with the initial framework used here for testing the Palaeoriver Hypothesis, our observations concerning the significant effect of pre-Pleistocene geological settlement and palaeorivers arrangement on biodiversity are in agreement with this recent finding. Finally, our findings propose new perspectives on the biogeography of Sundaland freshwater biotas and open new questions about the interplay between geology and palaeoecology during dispersal and colonization of Sundaland islands. Riverine organisms have constrained dispersal routes, whereas terrestrial organisms do not. As such, our studies warrant new researches on the biogeography of terrestrial rainforest biotas of Sundaland.

Acknowledgements

This paper is part of the doctoral study of AS, funded by a full scholarship of the Indonesian Endowment Fund for Education (LPDP) scholarship from the Indonesian Ministry of Finance. The authors thank Hari Sutrisno and Cahyo Rahmadi as well as the staff of the Zoology division of the Research Centre for Biology (Indonesian Institute of Sciences) for hosting AS during the first year of her PhD. We thank David J. Lohman and one anonymous reviewer for the relevant comments on earlier versions of the manuscript. This publication is ISEM publication number SUD-xxx.

Chapter 2
Disentangling the Taxonomy of the Subfamily Rasborinae (Cypriniformes, Danionidae) in Sundaland Using DNA Barcodes

*Arni Sholihah[1,2], Erwan Delrieu-Trottin[2,3], Tedjo Sukmono[4], Hadi Dahruddin[2,5], Renny Risdawati[6], Roza Elvyra[7], Arif Wibowo[8,9], Kustiati Kustiati[10], Frédéric Busson[2,11], Sopian Sauri[5], Ujang Nurhaman[5], Edmond Dounia[12], Muhamad Syamsul Arifin Zein[5], Yuli Fitriana[5], Ilham Vemendra Utama[5], Zainal Abidin Muchlisin[13], Jean-François Agnèse[2], Robert Hanner[14], Daisy Wowor[5], Dirk Steinke[14], Philippe Keith[11], Lukas Rüber[15,16] & Nicolas Hubert[2]**

1 Institut Teknologi Bandung, School of Life Sciences and Technology, Bandung, Indonesia.
2 UMR 5554 ISEM (IRD, UM, CNRS, EPHE), Université de Montpellier, Place Eugène Bataillon, 34095, Montpellier cedex 05, France.
3 Museum für Naturkunde, Leibniz-Institut für Evolutions und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Invalidenstrasse 43, Berlin, 10115, Germany.
4 Universitas Jambi, Department of Biology, Jalan Lintas Jambi - Muara Bulian KM 15, 36122, Jambi, Sumatra, Indonesia.
5 Division of Zoology, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Jalan Raya Jakarta Bogor KM 46, Cibinong, 16911, Indonesia.
6 Department of Biology Education, STKIP PGRI Sumatera Barat, Jl Gunung Pangilun, Padang, 25137, Indonesia.
7 Universitas Riau, Department of Biology, Simpang Baru, Tampan, Pekanbaru, 28293, Indonesia.
8 Southeast Asian Fisheries Development Center, Inland Fisheries Resources Development and Management Department, 8 Ulu, Seberang Ulu I, Palembang, 30267, Indonesia.
9 Research Institute for Inland Fisheries and Fisheries extensions, Agency for Marine and Fisheries Research, Ministry of Marine Affairs and Fisheries., Jl. H.A. Bastari No. 08, Jakabaring, Palembang, 30267, Indonesia.
10 Universitas Tanjungpura, Department of Biology, Jalan Prof. Dr. H. Hadari Nawawi, Pontianak, 78124, Indonesia.
11 UMR 7208 BOREA (MNHN-CNRS-UPMC-IRD-UCBN), Muséum National d'Histoire Naturelle, 43 rue Cuvier, 75231, Paris cedex 05, France.
12 UMR 5175 CEFE (IRD, UM, CNRS, EPHE), 1919 route de Mende, 34293, Montpellier cedex 05, France.
13 Syiah Kuala University, Faculty of Marine and Fisheries, Banda, Aceh, 23111, Indonesia.
14 Department of Integrative Biology, Centre for Biodiversity Genomics, 50 Stone Rd E, Guelph, ON, N1G2W1, Canada.
15 Naturhistorisches Museum Bern, Bernastrasse 15, Bern, 3005, Switzerland.
16 Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, 3012, Bern, Switzerland.
*email: nicolas.hubert@ird.fr

Scientific Reports volume 10, article number: 2818 (2020)
https://doi.org/10.1038/s41598-020-59544-9
Received: 21 November 2019; Accepted: 20 January 2020;
Published online: 18 February 2020

Abstract

Sundaland constitutes one of the largest and most threatened biodiversity hotspots; however, our understanding of its biodiversity is afflicted by knowledge gaps in taxonomy and distribution patterns. The subfamily Rasborinae is the most diversified group of freshwater fishes in Sundaland. Uncertainties in their taxonomy and systematics have constrained its use as a model in evolutionary studies. Here, we established a DNA barcode reference library of the Rasborinae in Sundaland to examine species boundaries and range distributions through DNA-based species delimitation methods. A checklist of the Rasborinae of Sundaland was compiled based on online catalogs and used to estimate the taxonomic coverage of the present study. We generated a total of 991 DNA barcodes from 189 sampling sites in Sundaland. Together with 106 previously published sequences, we subsequently assembled a reference library of 1097 sequences that covers 65 taxa, including 61 of the 79 known Rasborinae species of Sundaland. Our library indicates that Rasborinae species are defined by distinct molecular lineages that are captured by species delimitation methods. A large overlap between intraspecific and interspecific genetic distance is observed that can be explained by the large amounts of cryptic diversity as evidenced by the 166 Operational Taxonomic Units detected. Implications for the evolutionary dynamics of species diversification are discussed.

Introduction

Over the past two decades, the spectacular aggregation of species in biodiversity hotspots has attracted attention by scientists and stakeholders alike (Myers et al. 2000; Lamoreux et al. 2006; Schipper et al. 2008; Hoffmann et al. 2010). However, this exceptional concentration of often-endemic species at small spatial scales is threatened by the rise of anthropogenic disturbances. Of the 26 initially identified terrestrial biodiversity hotspots (Myers et al. 2000), the ones located in Southeast Asia (SEA) (Indo-Burma, Philippines, Sundaland and Wallacea) currently rank among the most important both in terms of species richness and the extend of endemism but also rank as the most threatened by human activities (Hoffmann et al. 2010). Sundaland is currently the most diverse terrestrial biodiversity hotspot of SEA and is the most threatened (Lohman et al. 2011). Sundaland comprises Peninsular Malaysia and the islands of Java, Sumatra, Borneo, and Bali and its diversity originates from the complex geological history of the region, linked to major tectonic changes in the distribution of land and sea during the last 50 Million years (My) (Hall 2012), but also from eustatic fluctuations that have sporadically connected and disconnected Sundaland landmasses during glacial-interglacial cycles in the Pleistocene (Voris 2000; Woodruff 2010; de Bruyn et al. 2014). Therefore, Sundaland biogeography has

received increased attention over the past decade resulting in the detection of contrasting spatial and temporal patterns in various groups (de Bruyn et al. 2013, 2014; Dong et al. 2018; Hendriks et al. 2019; O'Connell et al. 2019).

Species richness within vertebrate groups is high in Sundaland (Myers et al. 2000) and freshwater fishes are no exceptions to that. With more than 900 species reported to date, and with nearly 45 percent of endemism, Sundaland's ichthyofauna is the largest in SEA and accounts for nearly 75 percent of the entire ichthyodiversity of the Indonesian archipelago (Hubert et al. 2015b). The inventory of Sundaland's freshwater fishes started more than two centuries ago and despite the acceleration of species discovery over the last three decades, it is still a work in progress (Hubert et al. 2015b). The complexity of this inventory was partly exacerbated by the abundance of minute species *i.e*. less than 5 cm length (Hubert et al. 2015b), but also by the inconsistent use of species names through time for old descriptions due to the loss of type specimens or uncertainties in the location of type localities (Keith et al. 2017; Hubert et al. 2019). The family Cyprinidae *sensu lato* is a particularly good example for the complexity of Sundaland freshwater fishes taxonomy and systematics. The systematics of this large family of Cypriniformes, with over 3,000 species, has been controversial for more than a century (Conway et al. 2010). Based on recent molecular phylogenetic studies (Tang et al. 2010; Stout et al. 2016; Hirt et al. 2017), Tan and Armbruster (2018) proposed a new classification dividing the Cyprinidae *sensu lato* into 12 families. The subfamily Rasborinae (Cypriniformes, Cyprinoidei, Danionidae) comprises roughly 140 species in 11 genera: *Amblypharyngodon*, *Boraras*, *Brevibora*, *Horadandia*, *Kottelatia*, *Pectenocypris*, *Rasbora*, *Rasboroides*, *Rasbosoma*, *Trigonopoma*, and *Trigonostigma* (Tan and Armbruster 2018). In Sundaland the subfamily Rasborinae is represented by 79 species in 7 genera. The genera *Amblypharyngodon*, *Horadandia*, *Rasboroides*, and *Rasbosoma* do not occur in Sundaland. By far the most species rich rasborine genus is *Rasbora* with over 100 species in total and 65 species in Sundaland. Long considered a catch-all group, several attempts have been made to provide a classification of the genus *Rasbora* that reflects phylogeny. In a comprehensive revision, Brittan (1954) recognized 3 subgenera (*Rasbora*, *Rasboroides*, and *Megarasbora*) and divided *Rasbora* into 8 species complexes, now regarded as species groups (Liao et al. 2010) (Fig. 2.1). Subsequent authors have erected several new genera or suggested new species composition for the various *Rasbora* species groups (Kottelat and Vidthayanon 1993;

Kottelat and Witte 1999; Liao et al. 2010; Tang et al. 2010). Clearly, to better understand the evolutionary history of this unique group, the taxonomy and systematic of the Rasborinae needs to be better understood.

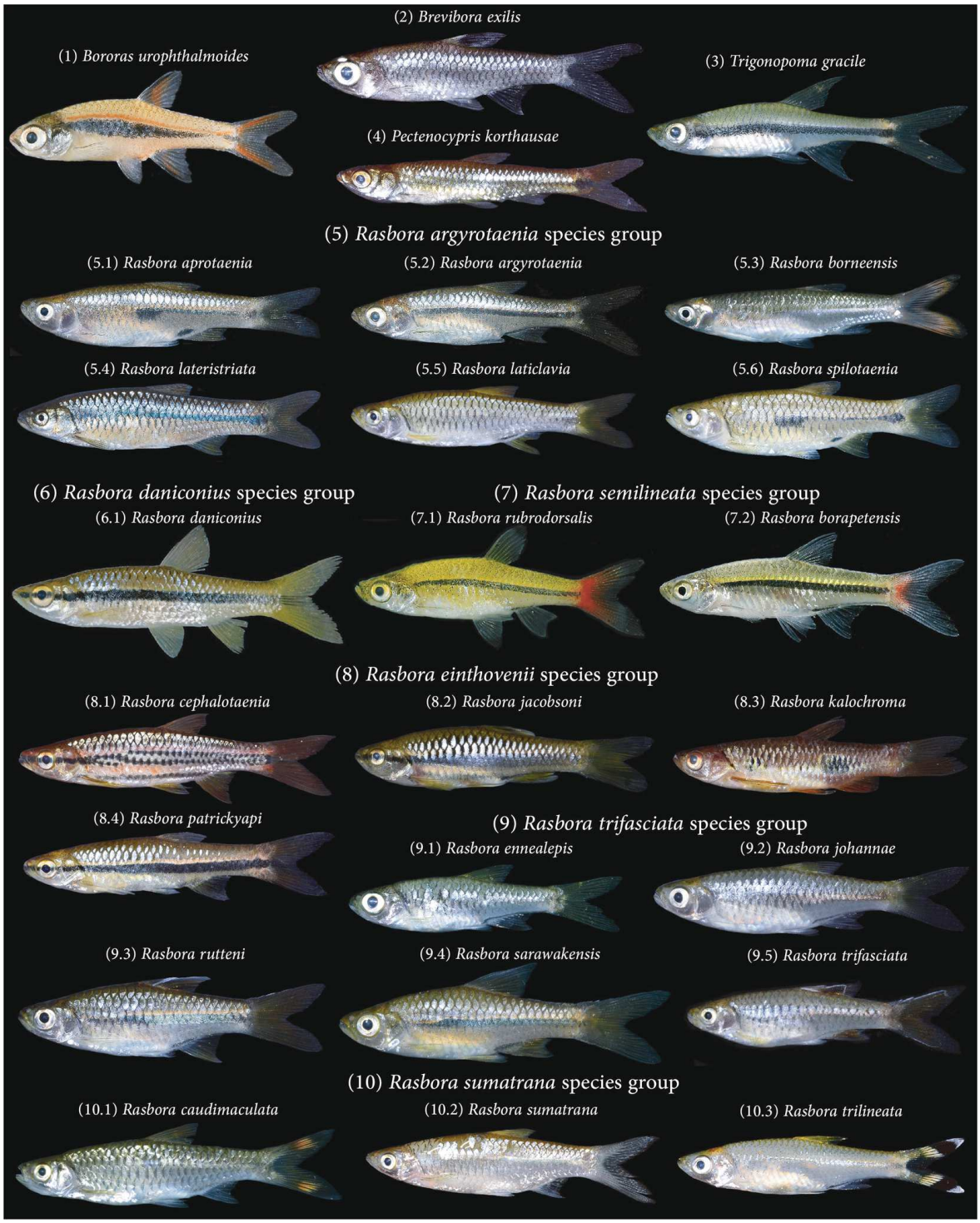

Figure 2.1 Selected species of Rasborinae that illustrate the diversity of the subfamily in Sundaland. All pictures, except 1, 6.1, 7.1 and 7.2, originate from the Barcode of Life Datasystem (dx.doi.org/110.5883/DS-BIFRA, Creative Commons Attribution - Non Commercial - Share Alike), pictures 1, 6.1 and 7.2 originate from FFish.asia (https://ffish.asia, Creative Commons Attribution - Non Commercial - Share Alike).

The use of standardized DNA-based approaches to the inventory of Sundaland's ichthyofauna resulted in the detection of considerable knowledge gaps (Dahruddin et al. 2017; Keith et al. 2017; Hubert et al. 2019). In addition, substantial levels of cryptic diversity (*i.e*. morphologically unrecognized diversity) were repeatedly reported for a wide range of Sundaland freshwater fish taxa (Nguyen et al. 2008; de Bruyn et al. 2013; Lim et al. 2016a; Beck et al. 2017; Conte-Grand et al. 2017; Dahruddin et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018) including the Rasborinae (Hubert et al. 2019). The taxonomy of most Rasborinae species, particularly so for the genus *Rasbora*, remains challenging due to the diversity of the group and the morphological similarity of many closely related species. As a consequence, the actual distribution ranges of many species of Rasborinae are not well known.

This study aims to re-examine Rasborinae diversity in Sundaland. We generated a DNA barcode reference library to (1) explore biological species boundaries with DNA-based species delimitation methods, (2) validate species identity, taxonomy and precise range distribution by producing DNA barcodes from type localities or neighboring watersheds, (3) validate or revise of the previously published DNA barcodes records for the subfamily Rasborinae available on GenBank.

Materials and Methods

Sampling and collection management

Material used in the present study is the result of a collective effort to assemble a global Rasborinae DNA barcode reference library through various field sampling efforts conducted by several of the coauthors in Sundaland over the past decade. Specimens were captured using gears such as electrofishing, seine nets, cast nets and gill nets across sites that encompass the diversity of freshwater lentic and lotic habitats in Sundaland (Fig. 2.2). Specimens were identified following original descriptions where available, as well as monographs (Kottelat et al. 1993; Kottelat 2013). Species names were further validated using several online catalogs (Froese and Pauly 2014; Eschmeyer et al. 2018). Specimens were photographed, individually labeled and voucher specimens were preserved in a 5% formalin solution. Prior to fixation a fin clip or a muscle biopsy was taken and fixed separately in a 96% ethanol solution for further genetic analyses. Both tissues and voucher specimens were

deposited in the national collections at the Museum Zoologicum Bogoriense (MZB), Research Center for Biology (RCB), Indonesian Institute of Sciences (LIPI).

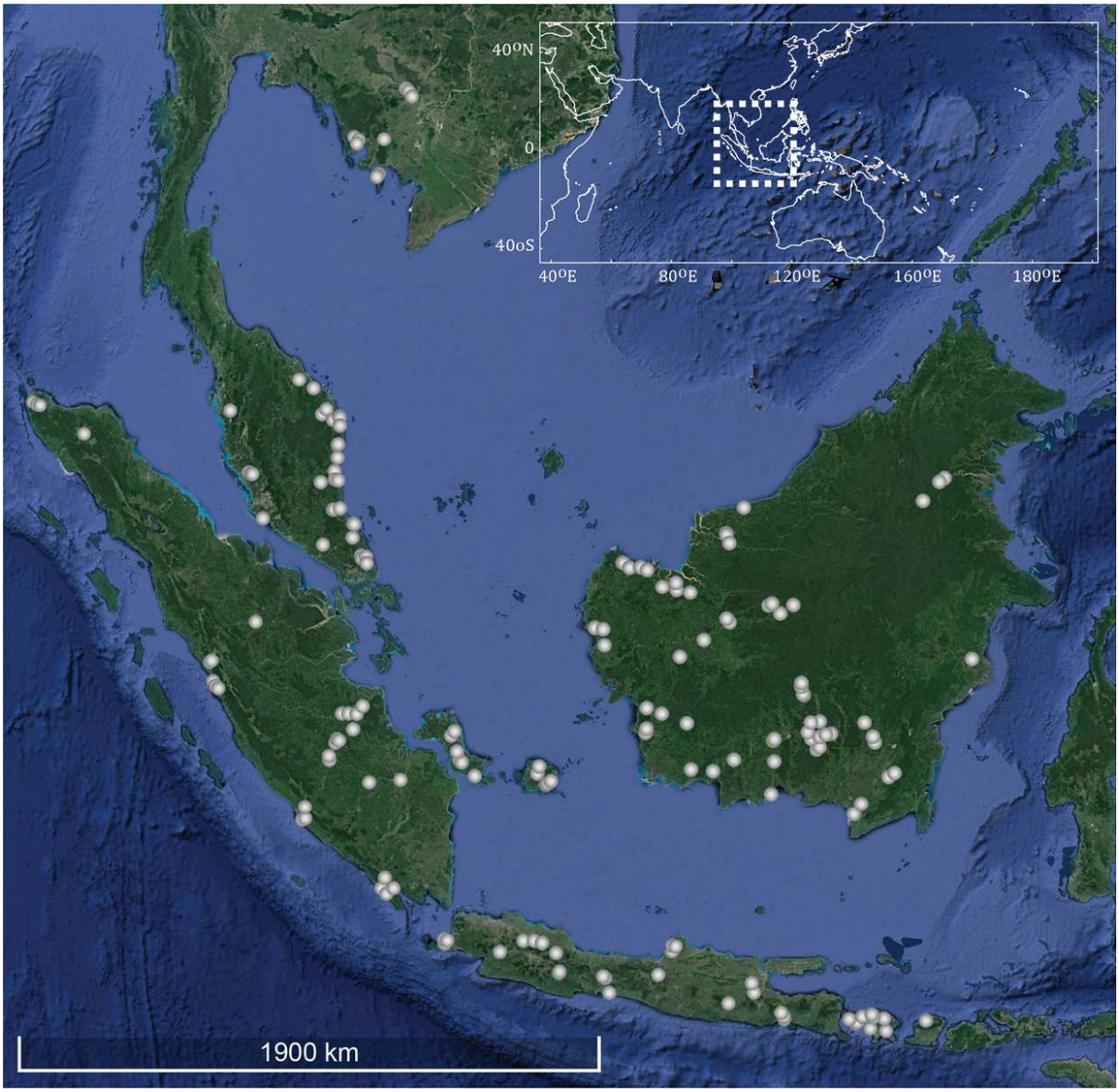

Figure 2.2 Collection sites for the newly generated 991 samples analyzed here. Each dot may represent several collection sites. Map data: Google, DigitalGlobe. Modified using Adobe Illustrator CS5 v 15.0.2. http://www.adobe.com/products/illustrator.html.

Assembling a checklist of the Sundaland Rasborinae

A checklist of the Rasborinae species occurring in Sundaland was assembled from available online catalogs including Fishbase (Froese and Pauly 2014) and Eschmeyer's Catalog of Fishes (Eschmeyer et al. 2018) as detailed in Hubert et al. (2015b). This checklist was used to estimate the taxonomic coverage of the present DNA barcoding campaign and to identify type localities for each species. The following

information was included: (1) authors of the original description, (2) type locality, (3) latitude and longitude of the type locality, (4) holotype and paratypes catalog numbers, (5) distribution in Sundaland. This information is available as online Supplementary Material (Table S2.1).

Sequencing and international repositories

Genomic DNA was extracted using a Qiagen DNeasy 96 tissue extraction kit following manufacturer's specifications. A 651-bp segment from the 5′ region of the cytochrome oxidase I gene (COI) was amplified using primers cocktails C_FishF1t1/C_FishR1t1 including M13 tails52. PCR amplifications were done on a Veriti 96-well Fast (ABI-AppliedBiosystems) thermocycler with a final volume of 10.0 µl containing 5.0 µl Buffer 2× 3.3 µl ultrapure water, 1.0 µl each primer (10 µM), 0.2 µl enzyme Phire Hot Start II DNA polymerase (5U) and 0.5 µl of DNA template (~50 ng). Amplifications were conducted as followed: initial denaturation at 98 °C for 5 min followed by 30 cycles denaturation at 98 °C for 5 s, annealing at 56 °C for 20 s and extension at 72 °C for 30 s, followed by a final extension step at 72 °C for 5 min. The PCR products were purified with ExoSap-IT (USB Corporation, Cleveland, OH, USA) and sequenced in both directions. Sequencing reactions were performed using the "BigDye Terminator v3.1 Cycle Sequencing Ready Reaction" and sequencing was performed on the automatic sequencer ABI 3130 DNA Analyzer (Applied Biosystems). DNA barcodes obtained at the Naturhistorisches Museum Bern were generated as previously described in Conte-Grand et al. (2017).

The sequences and associated information were deposited on BOLD (Ratnasingham and Hebert 2007) and are available in the data set DS-BIFRA (Table S2.2, dx.doi.org/10.5883/DS-BIFRA). DNA sequences were submitted to GenBank (accession numbers are accessible directly at the individual records in BOLD). An additional set of 106 Rasborinae COI sequences were downloaded from GenBank (Table S2.3).

Genetic distances and species delimitation

Kimura 2-parameter (K2P) (Kimura 1980) pairwise genetic distances were calculated using the R package Ape 4.1 (Paradis et al. 2004). Maximum intraspecific and nearest neighbor genetic distances were calculated from the matrice of pairwise

K2P genetic distances using the R package Spider 1.5 (Brown et al. 2012). We checked for the presence of a barcoding gap, *i.e*. the lack of overlap between the distributions of the maximum intraspecific and the nearest neighbor genetic distances (Meyer and Paulay 2005), by plotting both distances and examining their relationships on an individual basis instead of comparing both distributions independently (Blagoev et al. 2016). A neighbor-joining (NJ) tree was built based on K2P distances using PAUP 4.0a (Swofford 2001) in order to visually inspect genetic distances and DNA barcode clusters (Fig. S2.1). This NJ tree was rooted using *Sundadanio retiarius*.

Several alternative methods have been proposed for delimitating molecular lineages (Pons et al. 2006; Puillandre et al. 2012; Ratnasingham and Hebert 2013; Zhang et al. 2013). Each of these methods have pitfalls, particularly when it comes to singletons (*i.e*. delimitated lineages represented by a single sequence) and a combination of different approaches is increasingly used to overcome potential pitfalls arising from uneven sampling (Kekkonen and Hebert 2014; Kekkonen et al. 2015; Blair and Bryson 2017; Hubert et al. 2019; Shen et al. 2019). We used four different sequence-based methods of species delimitation. For the sake of clarity, we refer to species identified based on morphological characters as species while species delimited using DNA sequences are referred to as Operational Taxonomic Unit (MOTU) (Avise 1994; Moritz 1994; Vogler and Desalle 1994). MOTUs were delimitated using the following algorithms: (1) Refined Single Linkage (RESL) as implemented in BOLD and used to generate Barcode Index Numbers (BIN) (Ratnasingham and Hebert 2013), (2) Automatic Barcode Gap Discovery (ABGD) (Puillandre et al. 2012), (3) Poisson Tree Process (PTP) in its multiple rates version (mPTP) as implemented in the stand-alone software mptp_0.2.3 (Zhang et al. 2013; Kapli et al. 2017), (4) General Mixed Yule-Coalescent (GMYC) in its multiple rate version (mGMYC) as implemented in the R package Splits 1.0–19 (Fujisawa and Barraclough 2013). RESL and ABGD used DNA alignments as input files while a ML tree was used for mPTP and a Bayesian Chronogram based on a strict-clock model using a 1.2% of genetic distance per million year (Bermingham et al. 1997) for mGMYC. The mPTP algorithm uses a phylogenetic tree as an input file, thus, a maximum likelihood (ML) tree was first reconstructed using RaxML (Stamatakis 2014) based on a GTR + Γ substitution model. Then, an ultrametric and fully resolved tree was reconstructed using the Bayesian approach implemented in BEAST 2.4.8 (Bouckaert et al. 2014). Two Markov chains of 50 millions each were ran independently using the Yule pure birth model tree prior, a strict-clock model and

a GTR + I + Γ substitution model. Trees were sampled every 10,000 states after an initial burnin period of 10 millions. Both runs were combined using LogCombiner 2.4.8 and the maximum credibility tree was constructed using TreeAnnotator 2.4.7 (Bouckaert et al. 2014). Identical haplotypes were pruned for further species delimitation analyses.

Results

Sequencing of the DNA barcode marker Cytochrome Oxidase 1 (COI) yielded a total of 991 new sequences (Table S2.2) from 189 sampling sites distributed across Sundaland (Fig. 2.2). Together with 106 DNA barcodes mined from GenBank and BOLD (Table S2.3), we assembled a DNA barcode reference library of 1,097 sequences from 65 taxa of Rasborinae and 1 taxon of Sundadanionidae (*Sundadanio retiarius*). The number of specimens analyzed per species ranged from 1 to 143, with an average of 14.6 sequences per species and only six species represented by a single sequence. The sequences ranged from 459 bp to 651 bp long, with 99 percent of the sequences being above 500 bp length, and no stop codons were detected, suggesting that all the sequences correspond to functional mitochondrial COI sequences. DNA barcodes for 61 of the 79 nominal species of Rasborinae reported from Sundaland were recovered (approximately 78%) corresponding to the 7 Rasborinae genera currently recognized (Table S2.1). The present study achieved a complete coverage at the species level for the genera *Boraras* (2 species), *Brevibora* (3 species), *Kottelatia* (1 species), *Trigonopoma* (2 species) and *Trigonostigma* (3 species). In turn, two out of the three *Pectenocypris* species (66%) and 48 out of the 65 *Rasbora* species (74%) currently recognized in Sundaland were collected (Table S2.1). Geographically, our dataset includes 86% of the Rasborinae of Borneo (38 out of 44), all the *Rasbora* species of Java (4 species) and 68% of the Rasborinae species of Sumatra (26 out of 38) were collected (Table S2.1). Finally, four undescribed taxa are highlighted, two taxa of *Rasbora* in Java, one taxon of *Trigonostigma* in Borneo (Table S2.2) and an additional *Rasbora* taxon, previously assigned to *R. paucisqualis* in the literature (Table S2.3).

Species delimitation analyses provided varying numbers of Operational Taxonomic Units (MOTUs) among methods (Fig. 2.3): 129 for PTP, 95 for mPTP, 178 for GMYC, 191 for mGMYC, 175 for ABGD and 146 for RESL (Table S2.3). Our

consensus delimitation scheme yielded 166 MOTUs, including 165 MOTUs for the 65 Rasborinae taxa, 2.5-fold more than by using morphological characters. The number of MOTUs observed within species ranged from two for 22 species to 11 for *Trigonostigma pauciperforata* (Table 2.1). Based on the results of the species delimitation analyses, a re-examination of the original species identity associated with 105 DNA barcodes mined from BOLD and GenBank revealed 13 cases of conflicts that likely originated from mis-identifications (Table S2.4). These concerned the genera *Boraras* (four records, two species), *Brevibora* (two records, two species) and *Rasbora* (seven records, six species). Along the same line, 12 uncertain identifications were revised for the genera *Rasbora* (10 records, five taxa) and *Trigonostigma* (two records, one taxa).

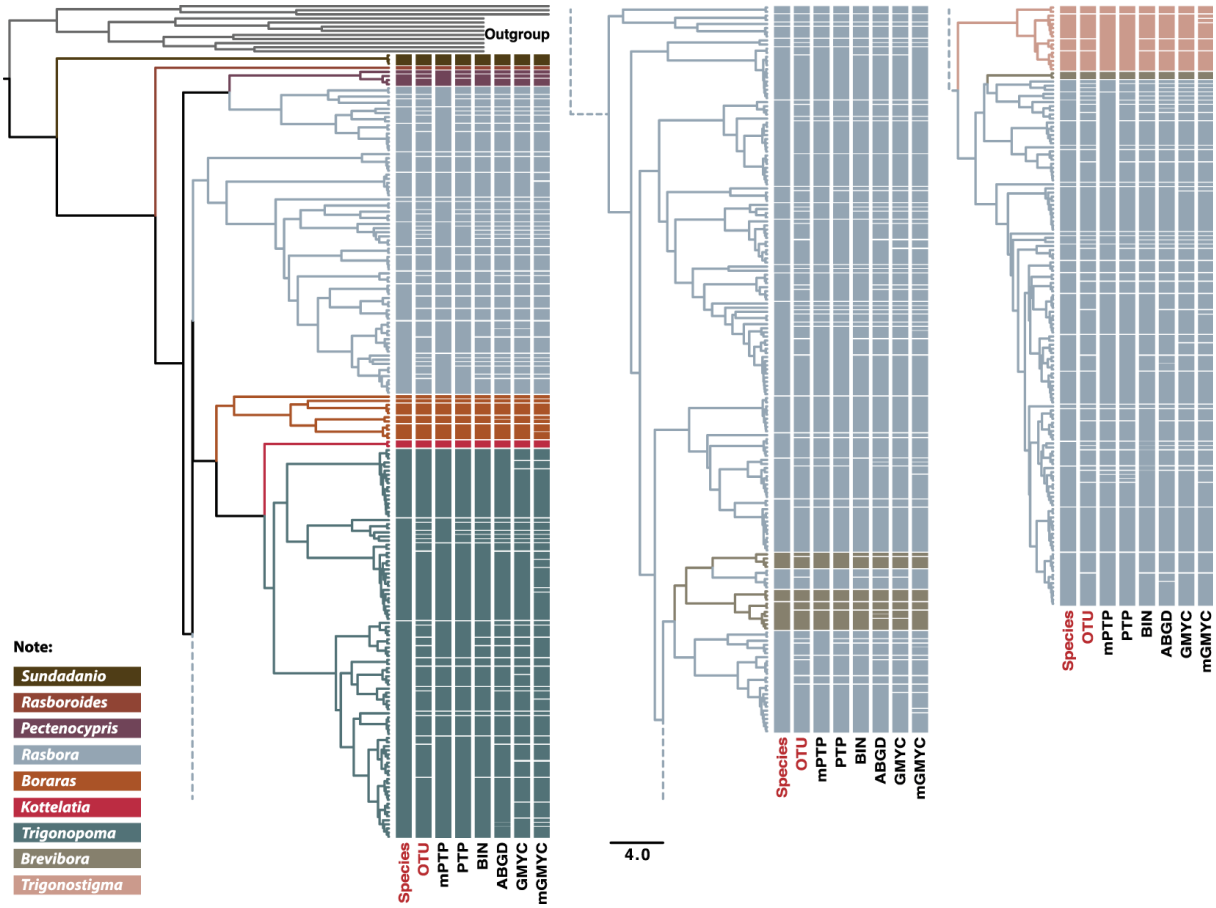

Figure 2.3 Bayesian maximum credibility tree of the Rasborinae DNA barcodes (identical haplotypes removed) and species delimitation according to GMYC, mGMYC, PTP, mPTP, ABGD, BIN and the 50% consensus delimitation.

Table 2.1 List of the morphological species displaying more than one MOTU including the maximum intraspecific and minimum nearest neighbor K2P distances for species and MOTUs

Species/MOTUs	Max. Intraspecfic Dist. (%)	Nearest Neighbor Dist. (%)	Species/MOTUs	Max. Intraspecfic Dist. (%)	Nearest Neighbor Dist. (%)
Brevibora cheya	4.29	5.99	*Rasbora einthovenii*	11.1	8.31
MOTU 105 (BOLD:AAY0408)	0	3.18	MOTU 73 (BOLD:ADY2667)	NA	7.45
MOTU 106 (BOLD:ADN0681)	1.3	3.18	MOTU 74 (BOLD:ADY1546)	0	7.45
Brevibora dorsiocellata	2.1	7.71	MOTU 75 (BOLD:ADY1017)	0	7.75
MOTU 102 (BOLD:ADY4509)	0	1.57	MOTU 77 (BOLD:ADW2748)	GenBank	GenBank
MOTU 103 (BOLD:ADN0680)	0.52	1.57	MOTU 78 (BOLD:ADN0813)	0	5.43
Rasbora aprotaenia	1.83	1.04	MOTU 79 (BOLD:AAU5112)	0	3.18
MOTU 140 (**BOLD:ADY6054**)	1.3	1.04	MOTU 80 (BOLD:ADO6360)	NA	2.1
OUT 139 (BOLD:ADZ0447)	NA	1.3	MOTU 81 (BOLD:ADY1549)	1.57	2.1
Rasbora argyrotaenia	5.67	5.97	MOTU 83 (BOLD:ADY0551)	0.52	1.3
MOTU 87 (BOLD:ADY7291)	0.26	5.1	MOTU 82 (BOLD:ADY0550)	0.78	1.3
MOTU 88 (BOLD:ACQ2593)	0.52	5.1	*Rasbora elegans*	1.57	1.04
Rasbora arundinata	2.63	2.1	MOTU 138 (**BOLD:ADY6054**)	0	1.04
MOTU 130 (BOLD:ADF6073)	0	2.1	MOTU 136 (BOLD:ADY7956)	NA	1.3
MOTU 131 (BOLD:ADN1040)	0	2.63	MOTU 137 (BOLD:ADZ0446)	0	1.3
Rasbora bankanensis	7.12	6.51	*Rasbora ennealepis*	9.14	6.51
MOTU 40 (BOLD:ACF1059)	GenBank	GenBank	MOTU 35 (BOLD:ADN3883)	0	5.94
MOTU 39 (BOLD:AAR2899)	0	1.3	MOTU 36 (BOLD:ADN3887)	0.78	3.97
MOTU 38 (BOLD:ADY4700)	NA	1.3	MOTU 37 (BOLD:ADY4386)	0.26	3.97
MOTU 41 (BOLD:ADY2504)	0	2.91	*Rasbora jacobsoni*	0	8.84
MOTU 42 (BOLD:ADY1545)	0	3.72	MOTU 66 (BOLD:ADW4597)	GenBank	GenBank
MOTU 43 (BOLD:ADY1544)	0	2.91	MOTU 67 (BOLD:ADN9402)	0	8.84
MOTU 44 (BOLD:ACC0430)	1.04	1.57	*Rasbora kalbarensis*	0.52	12.42
MOTU 144 (BOLD:ADY5341)	1.04	1.57	MOTU 20 (BOLD:AAY0409)	GenBank	GenBank
Rasbora beauforti	2.37	9.79	MOTU 21 (BOLD:ADN1457)	0.52	12.42
MOTU 33 (**BOLD:ADY4385**)	NA	2.1	*Rasbora kalochroma*	2.64	5.39
MOTU 34 (**BOLD:ADY4385**)	0.26	2.1	MOTU 71 (**BOLD:AAR2898**)	NA	1.3
Rasbora borapetensis	7.73	5.97	MOTU 72 (**BOLD:AAR2898**)	1.83	1.3
MOTU 86 (BOLD:ADY1548)	NA	7.43	*Rasbora kottelati*	2.37	5.39
MOTU 91 (BOLD:AAU5232)	0.52	5.97	MOTU 68 (BOLD:ADX8298)	GenBank	GenBank
Rasbora caudimaculata	1.83	7.71	MOTU 69 (BOLD:ADN0290)	0	2.1
MOTU 100 (BOLD:ADO5236)	0	1.83	MOTU 70 (BOLD:ADX9355)	0.26	2.1
MOTU 101 (BOLD:AAR2916)	NA	1.83	*Rasbora lateristriata*	1.83	2.9
Rasbora cephalotaenia	6.84	7.68	MOTU 141 (BOLD:ACQ7159)	1.3	1.57
MOTU 4 (BOLD:ADY2668)	2.36	5.41	MOTU 142 (BOLD:ACQ7160)	0	1.57
MOTU 5 (BOLD:AAI0355)	GenBank	GenBank	*Rasbora laticlavia*	4.55	4
MOTU 6 (BOLD:ADN8441)	0.26	3.17	MOTU 119 (BOLD:ADN8626)	NA	3.45
MOTU 7 (BOLD:AAI0356)	0.78	3.17	MOTU 120 (BOLD:ADO3612)	0.78	1.04
Rasbora daniconius	0.26	11.18	MOTU 121 (BOLD:ADY6696)	0.26	1.04
MOTU 2 (BOLD:ABX6594)	GenBank	GenBank	*Rasbora patrickyapi*	1.83	8.31
MOTU 3 (BOLD:ACA0514)	0.26	11.18	MOTU 146 (**BOLD:ADN2766**)	0	1.83
Rasbora dusonensis	1.57	10.73	MOTU 76 (**BOLD:ADN2766**)	0	1.57
MOTU 10 (BOLD:AAU2983)	0.26	1.3	MOTU 147 (**BOLD:ADN2766**)	NA	1.57
MOTU 9 (BOLD:ADN2767)	0	1.3			

Species/MOTUs	Max. Intraspecfic Dist. (%)	Nearest Neighbor Dist. (%)	Species/MOTUs	Max. Intraspecfic Dist. (%)	Nearest Neighbor Dist. (%)
Rasbora paucisqualis	5.97	7.63	*Rasbora tuberculata*	9.87	9.73
MOTU 26 (BOLD:ADY2665)	0	4.28	MOTU 24 (BOLD:ADN3886)	0	9.55
MOTU 27 (BOLD:ADX9120)	0	2.63	MOTU 25 (BOLD:ADN3884)	0.26	9.55
MOTU 28 (BOLD:ADY4316)	NA	1.57	*Rasbora vaillantii*	1.83	4
MOTU 29 (BOLD:ADY4315)	NA	1.57	MOTU 117 (BOLD:ADY8198)	0	1.83
MOTU 30 (BOLD:ADY4317)	0.26	1.83	MOTU 118 (BOLD:ADY8199)	0	1.83
Rasbora paviana	2.37	2.36	*Rasbora vulcanus*	0	7.69
MOTU 126 (**BOLD:AAD6182**)	0.52	1.83	MOTU 115 (BOLD:AAI0352)	GenBank	GenBank
MOTU 127 (**BOLD:AAD6182**)	GenBank	GenBank	MOTU 116 (BOLD:ADN3885)	0	7.69
MOTU 129 (BOLD:ADY6053)	1.3	1.83	*Trigonopoma gracile*	13.64	9.22
Rasbora rutteni	6.22	7.14	MOTU 46 (BOLD:ADN4644)	NA	6.26
MOTU 17 (BOLD:ADN4430)	1.04	5.93	MOTU 47 (BOLD:ADO0069)	0	1.57
MOTU 18 (BOLD:ADY4516)	0	2.37	MOTU 48 (BOLD:ADY2669)	NA	2.36
MOTU 19 (BOLD:ADN7331)	0	2.37	MOTU 49 (BOLD:ADY4282)	NA	1.57
*Rasbora sp.*1	2.63	3.45	MOTU 50 (BOLD:ADY6176)	0	1.57
MOTU 124 (BOLD:ACQ2698)	0	2.63	MOTU 145 (**BOLD:ACC0899**)	0.52	2.1
MOTU 125 (BOLD:ACQ2594)	0.52	2.63	MOTU 51 (**BOLD:ACC0899**)	2.1	2.1
Rasbora subtilis	3.99	7.06	*Trigonopoma pauciperforatum*	9.25	9.22
MOTU 111 (BOLD:ADN7332)	NA	3.99	MOTU 55 (BOLD:ADY1547)	1.83	1.83
MOTU 112 (BOLD:ADN3888)	1.57	3.99	MOTU 56 (BOLD:ADY1425)	0	1.83
Rasbora sumatrana	3.18	5.97	MOTU 57 (BOLD:ADY5548)	0.52	2.1
MOTU 89 (**BOLD:AAY0407**)	1.04	2.37	MOTU 58 (BOLD:AAV7972)	NA	4.81
MOTU 90 (**BOLD:AAY0407**)	0.78	2.37	MOTU 59 (BOLD:ACC0580)	1.3	4.54
Rasbora tornieri	1.57	9.51	MOTU 60 (**BOLD:ADY2666**)	0	1.3
MOTU 84 (**BOLD:ADL5624**)	NA	1.57	MOTU 61 (**BOLD:ADY2666**)	1.57	1.3
MOTU 85 (**BOLD:ADL5624**)	0	1.57	MOTU 62 (BOLD:ADN4643)	NA	3.45
Rasbora trilineata	10.97	7.69	MOTU 63 (BOLD:ACC0669)	0	3.99
MOTU 96 (BOLD:AAE7383)	1.83	2.36	MOTU 64 (BOLD:ADV1540)	2.37	1.83
MOTU 97 (BOLD:ADN9095)	0	5.96	MOTU 65 (BOLD:AAY0427)	1.83	1.83
MOTU 98 (BOLD:ADN7260)	NA	3.74	*Trigonostigma heteromorpha*	2.37	2.64
MOTU 99 (BOLD:ADN9096)	0	3.74	MOTU 108 (BOLD:AAJ8936)	0.26	1.83
MOTU 93 (**BOLD:AAE7384**)	GenBank	GenBank	MOTU 110 (BOLD:ABZ6147)	0.26	1.83
MOTU 94 (**BOLD:AAE7384**)	0	4.27			
MOTU 95 (BOLD:ADY1696)	0.26	2.36			

The examination of the maximum K2P genetic distances for species with multiple MOTUs and within MOTUs revealed large differences with maximum K2P distances ranging between 0.26 and 13.64 within species and between 0.00 and 2.37 within MOTUs (Table 2.1). This trend was largely confirmed by the distribution of the genetic distances at both species and MOTUs levels (Fig. 2.4). At the species level, the distribution of the maximum intraspecific K2P genetic distance broadly overlap with the distribution of the K2P distances to the nearest neighbor (Figs. 2.4a,b, Table S2.5) and no barcoding gap is observed. On average, the nearest neighbor K2P genetic distances are only 3.5-fold higher than the maximum intraspecific K2P distances. Plotting genetic distance for each species provides little improvement as a substantial number of species display maximum intraspecific K2P genetic distances higher than

the minimum distance to the nearest neighbor (Fig. 2.4c). At the MOTU level, the overlap is drastically reduced peaking between 0 and 0.99 for maximum intraspecific K2P distances and ranging between 1.0 and 1.99 for the K2P distance to the nearest neighbor (Fig. 2.4d,e). The distribution width of the maximum intraspecific K2P distance is much more restricted for MOTUs than species and fewer cases of maximum intraspe- cific distance higher than the minimum distance to the nearest neighbor are observed (Fig. 2.4f). At the MOTU level, the nearest neighbor K2P genetic distances were 7.2-fold higher on average than the maximum intraspecific K2P genetic distances.

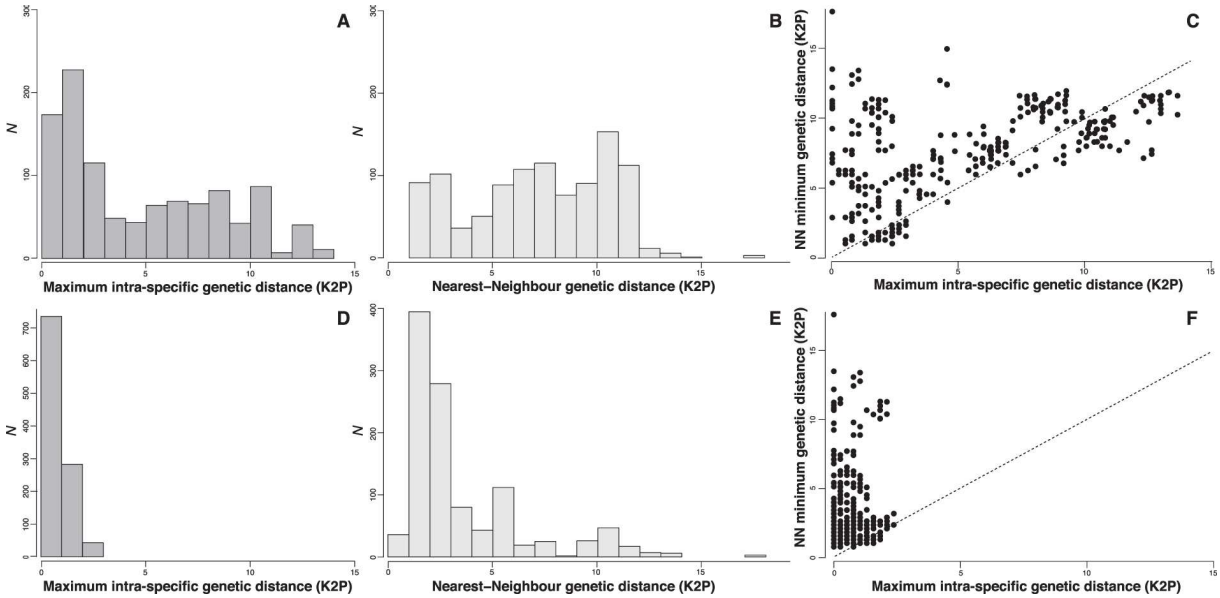

Figure 2.4 Summary of the distribution of the K2P distances.
(A,D) Maximum intraspecific K2P distances; (B,E) Minimum nearest-neighbor K2P distances; (C,F) Individual plotting of maximum intraspecific K2P distances and minimum nearest neighbor K2P distances. (A–C) Distributions of K2P distances for species delimitated using morphological characters. (D–F) Distributions of K2P distances for MOTUs delimitated by the 50% consensus among species delimitation methods.

Range distributions inferred from the new records generated for this study indicate that most type localities are embedded in the observed species range (Fig. 2.5). The degree of overlap between species range, however, largely varies among genera with little or no overlap observed for *Boraras*, *Pectenocypris*, *Trigonostigma* and most *Rasbora* species while a substantial amount of overlap is observed for *Brevibora* and *Trigonopoma* species (Fig. 2.5).

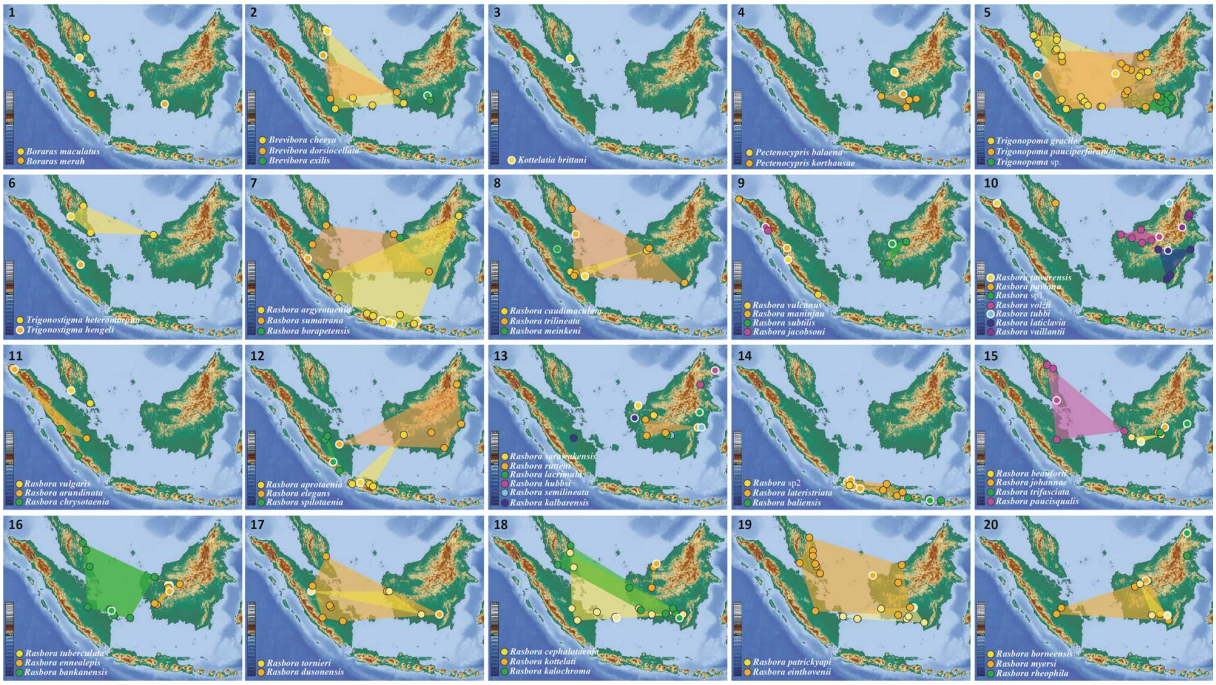

Figure 2.5 Maps depicting species distribution ranges as established based on the present sampling sites (black margin) and type localities (white margin) following the checklist generated for this study (Table S2.1).

1 Sampling sites and type localities of *Boraras maculatus* and *B. merah*. **2** Sampling sites, type localities and distribution ranges of *Brevibora cheeya*, *B. dorsiocellata* and *B. exilis*. **3** Type locality of *Kottelatia brittani*, sampling sites unknown, sequences originating from GenBank. **4** Sampling sites, type localities and distribution ranges of *Pectenocypris korthausea* and *P. balaena*. **5** Sampling sites, type localities and distribution ranges of *Trigonopoma gracile*, *T. pauciperforatum* and *T.* sp. (Hubert et al. 2019) **6** Sampling sites, type localities and distribution ranges of *Trigonostigma heteromorpha*, *T. hengeli* (sampling sites outside the map); *T. espei* not displayed, type locality outside the map and sampling sites unknown, sequences originating from GenBank. **7** Sampling sites, type localities and distribution ranges of *Rasbora argyrotaenia*, *R. sumatrana* and *R. borapetensis*, multiple type localities for *Rasbora argyrotaenia* as detailed in (Hubert et al. 2019), Type locality of *R. borapetensis* outside the map. **8** Sampling sites, type localities and distribution ranges of *Rasbora caudimaculata*, *R. trilineata* and *R. meinkeni*, sampling sites of *R. meinkeni* unknown, sequence originating from GenBank. **9** Sampling sites, type localities and distribution ranges of *Rasbora vulcanus*, *R. maninjau*, *R. subtilis* and *R. jacobsoni*. **10** Sampling sites, type localities and distribution ranges of *Rasbora tawarensis*, *R. paviana*, R. sp. 1 (Hubert et al. 2019), *R. volzii*, *R. tubbi*, *R. laticlavia* and *R. vaillanti*, sampling sites corresponding to the type locality for *Rasbora tawarensis*. **11** Sampling sites, type localities and distribution ranges of *Rasbora vulgaris*, *R. arundinata* and *R. chrysotaenia*, type locality of *Rasbora chrysotaenia* located in Sumatra with no further details (Table S2.1). **12** Sampling sites, type localities and distribution ranges of *Rasbora aprotaenia*, *R. elegans* and *R. spilotaenia*. **13** Sampling sites, type localities and distribution ranges of *Rasbora sarawakensis*, *R. rutteni*, *R. lacrimula*, *R. hubbsi*, *R. semilineata* and *R. kalbarensis*. **14** Sampling sites, type localities and distribution ranges of *Rasbora* sp. 2 (Hubert et al. 2019), *R. lateristriata* and *R. baliensis*. **15** Sampling sites, type localities and distribution ranges of *Rasbora beauforti*, *R. johannae*, *R. trifasciata* and *R. paucisqualis*. **16** Sampling sites, type localities and distribution ranges of *Rasbora tuberculata*, *R. ennealepis* and *R. bankanensis*. **17** Sampling sites, type localities and distribution ranges of *Rasbora tornieri* and *R. dusonensis*. **18** Sampling sites, type localities and distribution ranges of *Rasbora cephalotaenia*, *R. kottelati* and *R. kalochroma*. **19** Sampling sites, type localities and distribution ranges of *Rasbora patrickyapi* and *R. einthovenii*. **20** Sampling sites, type localities and distribution ranges of *Rasbora borneensis*, *R. myersi*, *R. rheophila*. Sampling sites and type locality of *R. daniconius* not displayed, outside the map. Each locality may represent several sampling sites. Map data: https://maps-for-free.com/. Modified using Adobe Illustrator CS5 v 15.0.2. http://www.adobe.com/products/illustrator.html.

Discussion

This study represents the most comprehensive molecular survey conducted for the subfamily Rasborinae (Tang et al. 2010; Collins et al. 2012). Our DNA barcode reference library consists of 65 Rasborinae species distributed across 7 genera and covering 78% of the Rasborinae diversity reported from Sundaland. DNA barcoding delivers reliable species-level identifications when taxa possess unique COI sequence clusters characterized by multiple private mutations. This condition was met for all the Rasborinae species examined here and no cases of retention of ancestral polymorphism were detected (Funk and Omland 2003). However, this clearly contrasts with multiples discrepancies observed within the set of previously published COI sequences obtained on GenBank and BOLD. About 25 percent of these records were either misidentified or associated with uncertain identifications. Such misidentifications were expected considering the morphological uniformity within some Rasborinae genera, particularly in the genus *Rasbora* where multiple cases of taxonomic conflicts have been highlighted already (Siebert 1997; Kottelat 2012; Muchlisin et al. 2012; Ng and Kottelat 2013; Hubert et al. 2019). Unexpectedly, most of the conflicts we detected were within the larger species of *Rasbora*, particularly those of the *Rasbora argyrotaenia* group and the *R. sumatra* group, and not within closely related smaller species such as members of the *R. trifasciata* group (Fig. 2.1). In facts, conflicts in species level population assignments have been previously reported for the *R. argyrotaenia* group in Java and Bali where *R. lateristriata* and *R. baliensis* have been confounded for decades as recently revealed by reexamination of species boundaries and distribution through DNA barcodes (Hubert et al. 2019). Other morphologically similar species of the *Rasbora argyrotaenia* group have been previously confused with *R. lateristriata*, such as *R. elegans*, *R. spilotaenia* and *R. chrysotaenia*. These species are difficult to separate due to overlapping meristic counts and coloration patterns (Kottelat et al. 1993). Our study, however, highlights that these species have disjunct range distributions (Fig. 2.5) and cluster into well-differentiated mitochondrial lineages (Fig. S2.1, Table S2.3). Several of the detected misidentifications also involve species from different *Rasbora* species groups (Liao et al. 2010) such as *Rasbora dusonensis*, from the *R. argyrotaenia* group, that has been previously mistaken for *R. sumatrana* from the *sumatrana* group and *R. myersi*, from the *R. sumatrana* group, that has been confounded with *R. dusonensis* from the *argyrotaenia* group. Despite being distantly

related (Fig. S2.1), these species show overlapping meristic counts and similar coloration patterns with no dark spots on the body (Kottelat et al. 1993). This result further calls for a broader assessment of the monophyly of the different *Rasbora* groups, previously identified by Liao et al. (2010), as they are poorly supported by our study (Fig. S2.1).

The observed average ratio of 3.5 between intraspecific and interspecific distances is very low compared to earlier values found for the Javanese ichtyofauna, where minimum nearest neighbor genetic distances are on average 28-fold higher than the maximum intraspecific genetic distances (Dahruddin et al. 2017). This value is also very low in comparison to previous large-scale fish DNA barcode surveys (Hubert et al. 2008, 2012, 2017; April et al. 2011; Pereira et al. 2013; Shen et al. 2019). This deviation can be attributed to a substantial amount of cryptic diversity revealed by our species delimitation analyses. For 61 species, delimitated on the basis of morphological characters, and validated by a match between species range distributions and type localities, we recovered a total of 166 MOTUs. When accounting for this cryptic diversity the ratio between the minimum nearest neighbor and maximum intraspecific distances rose to 7.5. Earlier large scale surveys in Sundaland already pointed to substantial levels of cryptic diversity (Lim et al. 2016a; Beck et al. 2017; Conte-Grand et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018) and it has also been demonstrated that small-size species are more sensitive to fragmentation, experience faster genetic drift and as such accumulate cryptic diversity at a faster rate than large-size species (April et al. 2013; Hubert et al. 2017). Along the same line, small-size species are more frequently confounded and lumped together, a bias that tend to inflate the proportion of hidden diversity (Kottelat et al. 2006).

We found very high numbers of MOTUs with deep genetic divergences (up to 13.64% in *Trigonopoma gracile*) in a number of species (ranging from 7 to 11) such as in *Rasbora bankanensis*, *Rasbora einthovenii*, *Rasbora trilineata*, *Trigonopoma gracile* and *Trigonopoma pauciperforatum*. These five species also display some of the widest range distributions in Sundaland with MOTUs occurring in Borneo, Sumatra, Peninsular Malaysia and several small islands across the Java sea (*R. bankanensis*, Fig. 2.5.16; *R. einthovenii*, Fig. 2.5.19; *R. trilineata*, Fig. 2.5.8; *T. gracile*, Fig. 2.5.5; *T. pauciperforatum*, Fig. 2.5.4). However, the scarcity of MOTU range overlap for those species suggests ongoing population fragmentation across the species range distribution (Tables S2.2 and S2.3). This pattern is likely connected to the complex

geological history of Sundaland which over the last 10 Million years was influenced by the subduction activity of the Asian and Australian plates and the resulting intense volcanic activity which produced multiple volcanic arches (Lohman et al. 2011). Furthermore, climatic fluctuations during the Pleistocene induced major sea levels changes leading to merging of Sundaland landmasses during glacial maxima and multiple fragmentations during glacial sea level low-stands (Voris 2000; Woodruff 2010). In such dynamic landscapes, complex patterns of distribution and high lineage diversity are to be expected (de Bruyn et al. 2013). The influence of eustatic fluctuations in Sundaland is exemplified by *Rasbora bankanensis*, *Rasbora einthovenii*, *Rasbora trilineata*, *Trigonopoma gracile* and *Trigonopoma pauciperforatum* all of which display wide range distributions among watersheds neighboring the Java sea. Those have been repeatedly connected during glacial maxima (Figs. 2.5.5, 5(8), 2.5.16 and 2.5.19). This pattern strongly contrasts with the lower genetic diversity and restricted range distribution of the species occurring in the Eastern part of Borneo such as *Rasbora vaillantii* (Fig. 2.5.10), *R. laticlavia* (Fig. 2.5.10), *R. trifasciata* (Fig. 2.5.15) and *R. reophila* (Fig. 2.5.20) or species occurring in the Western part of Sumatra such as *Rasbora vulcanus* (Fig. 2.5.9), *R. maninjau* (Fig. 2.5.9), *R. jacobsoni* (Fig. 2.5.9), *R. tawarensis* (Fig. 2.5.10); *R. chrysotaenia* (Fig. 2.5.11) and *R. arundinata* (Fig. 2.5.11) and species in Java and Bali such as *Rasbora sp.* 1 (Fig. 2.5.10), *R. sp*. 2 (Fig. 2.5.14), *R. lateristriata* (Fig. 2.5.14) and *R. baliensis* (Fig. 2.5.14). These parts of Borneo, Sumatra and partially Java were disconnected from the central region of Sundaland around the Java sea during the Pleistocene. This trend highlights the sensitive status of the endemic Rasborinae species in the peripheral areas of Sundaland due to their highly restricted distribution ranges. The present study also argues against translocation programs for the most widespread species, considering the high proportion of cryptic diversity, if species and MOTUs identity are not determined through DNA barcodes (Hutama et al. 2017; Hubert et al. 2019).

Conclusions

The subfamily Rasborinae is the most diverse freshwater fish group of Sundaland and therefore represents an excellent model to explore the evolutionary response of local freshwater biotas to a dynamic geological history and repeated

eustatic fluctuations. Affected by taxonomic confusions for decades, the genus *Rasbora* has been left aside of recent large-scale molecular studies aimed at exploring the diversification of aquatic biotas in Sundaland. Our comprehensive DNA barcode reference library for the subfamily enables further evolutionary studies on the diversification of the group, in particular within the genus *Rasbora*, which allowed us to trace evolutionary dynamics at the local scale in Sundaland (Hubert et al. 2019). The contrasting patterns of molecular diversity and species range distributions between Rasborinae species inhabiting the watersheds neighboring the Java sea and the species located on the Eastern part of Borneo call for a larger assessment of their dynamics of species proliferation based on broader genomic analyses. Clearly, future studies will also have to address the systematics of the Rasborinae as no evidence supporting the monophyly of *Rasbora* nor the different *Rasbora* species groups are detected here.

Acknowledgements

The authors wish to thank Siti Nuramaliati Prijono, Bambang Sunarko, Witjaksono, Mohammad Irham, Marlina Adriyani, Ruliyana Susanti, Rosichon Ubaidillah, the late Renny K. Hadiaty, Hari Sutrisno and Cahyo Rahmadi at Research Centre for Biology (RCB-LIPI) in Indonesia; Edmond Dounias, Jean-Paul Toutain, Robert Arfi and Valérie Verdier from the 'Institut de Recherche pour le Développement'; Joel Le Bail and Nicolas Gascoin at the French embassy in Jakarta for their continuous support. We also would like to thank Eleanor Adamson, Hendry Budianto, Tob Chann Aun, Pak Epang, Herman Ganatpathy, Sébastien Lavoué, Michael Lo, Hendry Michael, Joshua Siow, Heok Hui Tan, Elango Velautham, Norsham S. Yaakob, and Denis Yong for their help in the field. We are also particularly thankful to Sumanta at IRD Jakarta for his help during the field sampling in Indonesia. Part of the present study was funded by the Institut de Recherche pour le Développement (UMR226 ISE-M and IRD through incentive funds) to N.H.), the MNHN (UMR BOREA) to P.K., the French Ichthyological Society (SFI) to P.K., the Foundation de France to P.K., the French embassy in Jakarta to N.H., the Natural Environmental Research Council (NERC, NE/F003749/1) to L.R. and Ralf Britz; National Geographic (8509-08) to L.R. and North of England Zoological Society-Chester Zoo to L.R. The present study and all associated methods were carried out in accordance with relevant

guidelines and regulation of the Indonesian Ministry of Research and Technology (Indonesia), the Economic Planning Unit, Prime Minister's Department (Malaysia), the Forest Department Sarawak (Malaysia), the Vietnam National Museum of Nature (Vietnam) and the Inland Fisheries Research and Development Institute (Cambodia). Field sampling in Indonesia was conducted according to the research permits 097/SIP/FRP/SM/IV/2014 for Philippe Keith, 60/EXT/SIP/FRP/SM/XI/2014 for Frédéric Busson, 41/EXT/SIP/FRP/SM/VIII/2014 for Nicolas Hubert, 200/E5/E5.4/SIP/2019 for Erwan Delrieu-Trottin and, 1/TKPIPA/FRP/ SM/I/2011 and 3/TKPIPA/FRP/SM/III/2012 for Lukas Rüber. The Fieldwork in Peninsular Malaysia and Sarawak was conducted under permits issued by the Economic Planning Unit, Prime Minister's Department, Malaysia (UPE 40/200/19/2417 and UPE 40/200/19/2534) and the Forest Department Sarawak (NCCD.970.4.4[V]-43) and were obtained with the help of Norsham S. Yaakob (Forest Research Institute Malaysia, Kepong, Kuala Lumpur, Malaysia). Luong Van Hao and Pham Van Luc (Vietnam National Museum of Nature) helped with arranging research permits in Vietnam and So Nam (Inland Fisheries Research and Development Institute, IFReDI) helped with arranging research permits in Cambodia. All experimental protocols were approved by the Indonesian Ministry of Research and Technology (Indonesia), the Indonesian Institute of Sciences (Indonesia), the Forest Department Sarawak (Malaysia), Economic Planning Unit of the Prime Minister's Department (Malaysia), the Vietnam National Museum of Nature (Vietnam) and the Inland Fisheries Research and Development Institute (Cambodia). It is a great pleasure to thank Soraya Villalba for generating the DNA barcodes at the Naturhistorisches Museum Bern. Sequence analysis was aided by funding through the Canada First Research Excellence Fund as part of the University of Guelph Food from Thought program. We thank Paul Hebert, Alex Borisenko and Evgeny Zakharov as well as BOLD and CCDB staff at the Centre for Biodiversity Genomics, University of Guelph for their valuable support. This publication has the ISEM number 2019-293-SUD.

Author Contributions

L.R. and N.H. designed the study. A.S., T.S., H.D., R.R., R.E., A.W., K.K., F.B., S.S., U.N., E.D., I.V.U., Z.A.M., D.W., P.K., L.R. and N.H. conducted the field sampling. A.S., E.D.T., T.S., H.D., A.W., L.R. and N.H. performed the morphological

identifications. H.D., S.S., U.N., F.B. and N.H. curated the specimen collection. A.S., E.D.T., H.D., M.S.A.Z., Y.F., I.V.U., R.H., D.S., L.R. and N.H. conducted the laboratory work. A.S., E.D.T., H.D., F.B., N.H., R.H. and D.S. curated the DNA barcode records in BOLD. A.S., E.D.T., H.D., J.F.A., L.R. and N.H. analyzed the data. A.S., E.D.T., H.D., J.F.A., D.S., L.R. and N.H. wrote the initial manuscript and all authors commented and approved the final version of the manuscript.

Chapter 3
Limited dispersal and in situ diversification drive the evolutionary history of Rasborinae fishes in Sundaland

Arni Sholihah[1,2], Erwan Delrieu-Trottin[1,3], Tedjo Sukmono[4], Hadi Dahruddin[5], Juliette Pouzadoux[1,6], Marie-Ka Tilak[1], Yuli Fitriana[5], Jean-François Agnèse[1], Fabien Condamine[1], Daisy Wowor[5], Lukas Rüber[7,8], Nicolas Hubert[1]

[1] UMR 5554 ISEM (IRD, UM, CNRS, EPHE), Université de Montpellier, Place Eugène Bataillon, 34095, Montpellier cedex 05, France.
[2] Institut Teknologi Bandung, School of Life Sciences and Technology, Jalan Ganesha 10, Bandung 40132, Indonesia.
[3] Museum für Naturkunde, Leibniz-Institut für Evolutions und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Invalidenstrasse 43, Berlin, 10115, Germany.
[4] Universitas Jambi, Department of Biology, Jalan Lintas Jambi - Muara Bulian KM 15, 36122, Jambi, Sumatra, Indonesia.
[5] Division of Zoology, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Jalan Raya Jakarta Bogor KM 46, Cibinong, 16911, Indonesia.
[6] UMR 5244 IHPE (CNRS, IFREMER, UM, UPVD), Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier cedex 05, France
[7] Naturhistorisches Museum Bern, Bernastrasse 15, Bern 3005, Switzerland
[8] Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland

Abstract

Aim: We examine the relative impact of Sundaland geology since the Oligocene and of Pleistocene Climatic Fluctuations on the diversification of a species-rich subfamily of Cypriniformes fishes widely distributed in Southeast Asia, the Rasborinae. We specifically tested if variations in the extent of exposed lands and island connectivity during Pleistocene eustasy (the Palaeoriver hypothesis) induced bursts of diversification.

Location: Sundaland

Taxon: Rasborinae (Actinopterygii, Cypriniformes, Danionidae)

Methods: We aggregated 1,017 DNA barcodes and 79 mitogenomes to delineate Molecular Operational Taxonomic Units (MOTUs) and further reconstruct a time-calibrated phylogeny of Rasborinae. Ancestral area estimations were conducted using both island and palaeoriver partitioning to examine the impact of island connectivity during Pleistocene eustasy on dispersal. Temporal trends of diversification are explored through a model-based approach.

Results: The origin of Sundaland lineages is dated at 33.41 Ma and four major clades are identified, which initiated their diversification between 31.10 and 25.95 Ma. Borneo Island and North Sunda palaeoriver are identified as the source of Sundaland Rasborinae. Geographical patterns of speciation indicate that most speciation events occurred within islands and constant birth rate models of diversification are the most likely models for all clades.

Conclusions: The geographical and historical context of diversification in Rasborinae provides little support to the Palaeoriver Hypothesis. The onset of Borneo isolation from mainland Asia triggered the initial diversification of the group (31-26 Ma). The late colonization of Java and Sumatra occurred through an assortment of dispersal events, poorly explained by Pleistocene eustasy, and frequently followed by *in situ* diversification.

Keywords: DNA-based species delimitation, Historical biogeography, Mitogenomes, Pleistocene Climatic Fluctuations, Eustasy, Dispersal

Introduction

Sundaland has long attracted the attention of evolutionary biologists. From his observations in the 19th century, Alfred Wallace already pointed out the biological uniqueness of the "Indo-Malay Islands" compared to neighbouring continental Asia and Celebes (Wallace 1869). It is now acknowledged that Sundaland's diversity and endemism occur in an area where the geological history is intricate (Hall 2009, 2013; Lohman et al. 2011). Including the Malay Peninsula and the islands of Borneo, Sumatra, Java and Bali, Sundaland emerged during the Oligocene, ca. 30 million years ago (Ma), as a promontory at the southern end of Eurasia (Fig. 3.1a). Complex tectonic movements during the Miocene triggered the formation of Borneo between 20 and 10 Ma (Figs. 3.1b & 3.1c) and the subsequent emergence of Sumatra and Java between 10 and 5 Ma (Figs. 3.1c & 1d). Insular Sundaland remained partially connected until entering the Pliocene (5.33-2.58 Ma). Upon entering early Quaternary (2.58 Ma), the Pleistocene Climatic Fluctuations (PCFs) led to global variations of average temperature (Zachos et al. 2008; Westerhold et al. 2020) and sea levels between 2.58 and 0.012 Ma (Miller et al. 2005). Geology and Pleistocene eustasy interacted in Sundaland (Voris 2000; Sathiamurthy and Voris 2006; Woodruff 2010; Sarr et al. 2019; Husson et al. 2019). During Pleistocene glacial periods, sea level dropped between -60m and -120m and created connections between Sundaland's islands (Fig. 3.1d). These exposed lands likely had freshwater drainage systems that extended between modern islands, the Palaeoriver Hypothesis (Kottelat et al. 1993; Voris 2000). Four major palaeorivers occurred in Southeast Asia: (1) East Sunda, (2) North Sunda, (3) Malacca straits, and (4) Siam (Fig. 3.1d). These palaeorivers likely impacted the dispersal of Sundaland's freshwater biotas and their evolutionary history (de Bruyn et al. 2013).

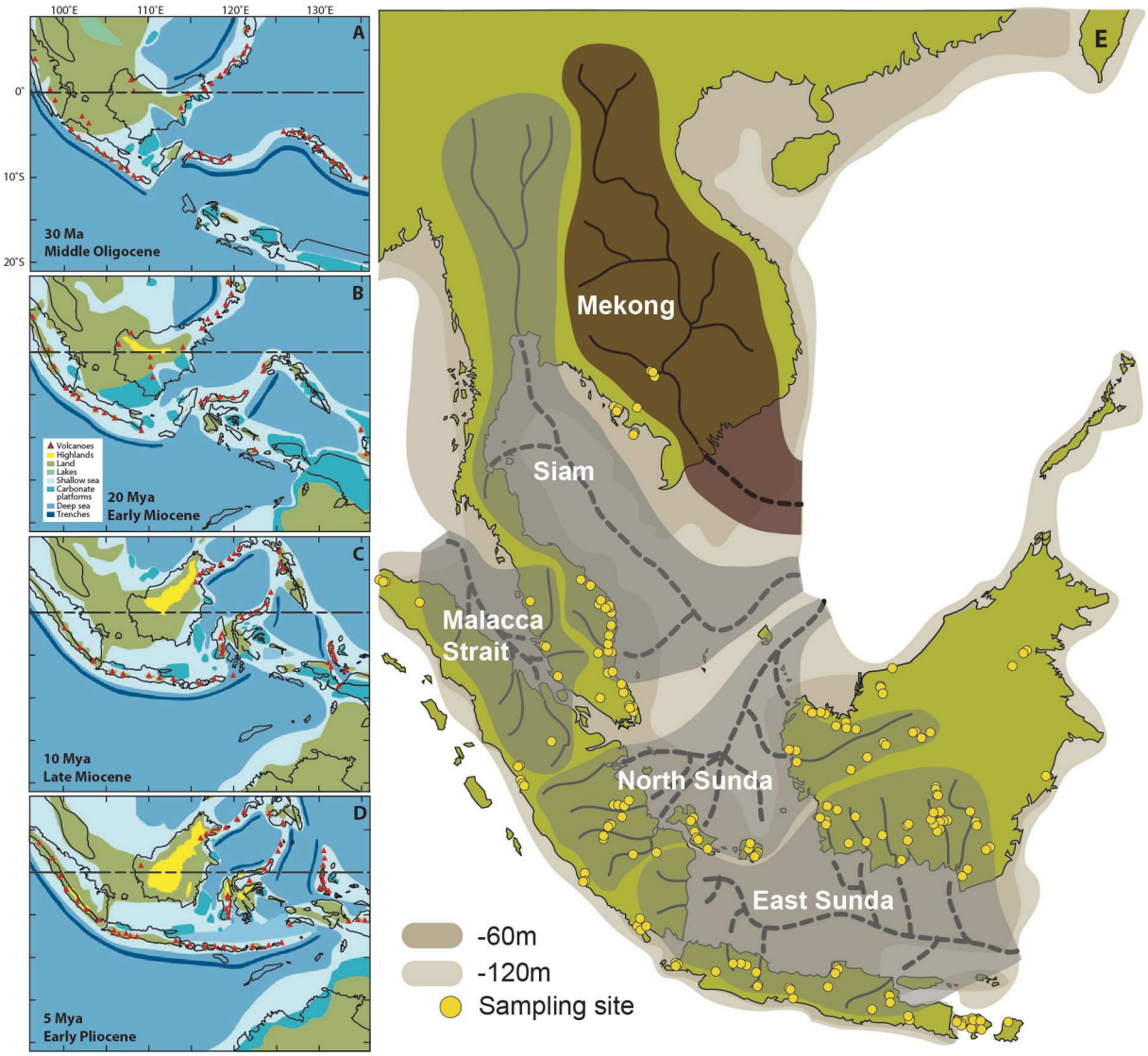

Figure 3.1 Geological reconstructions of the Indo-Australian archipelago since the middle Oligocene (modified from Lohman et al. 2011) and palaeoriver reconstruction in the Pleistocene (modified from Voris, 2000 and Woodruff, 2010).
A, middle Oligocene. B, early Miocene. C, late Miocene. D, early Pliocene. E, modern including limits of exposed land during -60 m and -120 m sea level drops, contour of the palaeoriver watersheds and sampling sites.

A high proportion of Sundaland contemporary freshwater diversity corresponds to cryptic lineages with Pleistocene origins (de Bruyn et al. 2014; Hubert et al. 2015b; Kusuma et al. 2016; Dahruddin et al. 2017; Nurul Farhana et al. 2018; Hubert et al. 2019; Sholihah et al. 2020). Several studies detected congruence between the distribution of some freshwater lineages and boundaries of palaeoriver watersheds (Dodson et al. 1995; Tan et al. 2012; de Bruyn et al. 2013; Beck et al. 2017). The temporal occurrence of numerous speciation events across clades during the Pleistocene suggests that PCFs triggered a burst of species proliferation in Sundaland through eustasy, a trend already suggested in other areas (Nores 1999; Barraclough

and Nee 2001; Wiens and Donoghue 2004; Hubert and Renno 2006; Mittelbach et al. 2007; Cannon et al. 2009; Condamine et al. 2015b). Even so, spatiotemporal dynamics of diversification in Sundaland cannot be explained solely by Pleistocene eustasy and others factors likely contributed to the build-up of Sundaland diversity including: (1) pre-Pleistocene geology (Dodson et al. 1995; Condamine et al. 2013b; de Bruyn et al. 2014; Beck et al. 2017), (2) dynamic interactions between insular and palaeoriver watershed boundaries (Esselstyn and Brown 2009; Brown et al. 2013; Papadopoulou and Knowles 2015a, 2015b), (3) varying dispersal abilities (Esselstyn et al. 2009; Pouyaud et al. 2009; Patel et al. 2011; Beck et al. 2017), and (4) Habitat rearrangements during PCFs (Heaney 1991; Bird et al. 2005; Wurster et al. 2019). By contrast to the Last Glacial Maximum (LGM) ca. 17,000 years ago, which housed the maximal extension of palaeoriver watersheds, Sundaland is currently considered as a refugia. This statement underlines the importance to understand the impact of PCFs on diversity patterns in Sundaland for effective conservation efforts. Sundaland is one of the most threatened biodiversity hotspots in Southeast Asia (Myers et al. 2000; Voris 2000; Sathiamurthy and Voris 2006; Cannon et al. 2009; Woodruff 2010; Mittermeier et al. 2011; Lohman et al. 2011; Hubert et al. 2015b).

Freshwater fishes are tightly dependent of watershed dynamics and constitute model systems to trace historical watershed dynamics (Bernatchez and Wilson 1998; Durand et al. 1999; Hubert et al. 2007; de Bruyn et al. 2013). Sundaland host several species-rich groups that constitute models to explore consequences of PCFs on freshwater diversity patterns (Hubert et al. 2015b). One of these groups is Rasborinae (Cypriniformes, Danionidae), a subfamily of iconic and highly diversified, small-size species with widespread distribution in Sundaland (Brittan 1972; Liao et al. 2011; Kusuma et al. 2016; Dahruddin et al. 2017; Tan and Armbruster 2018; Hubert et al. 2019; Sholihah et al. 2020). The Rasborinae comprises eleven genera varyingly distributed in Asia, of which seven are endemic of Sundaland or much more diverse here than in adjacent areas (Sholihah et al. 2020). These include *Boraras*, *Brevibora*, *Kottelatia*, *Pectenocypris*, *Rasbora*, *Trigonopoma* and *Trigonostigma* (Tan and Armbruster 2018; Sholihah et al. 2020). The systematic of the subfamily is still confused due to the lack of robust phylogenetic hypothesis of relationships within the group (Brittan 1972; Kottelat and Vidthayanon 1993; Kottelat and Witte 1999; Liao et al. 2010, 2011; Tan 2020), despite its monophyly is well supported (Saitoh et al. 2006; Tang et al. 2010; Stout et al. 2016; Tan and Armbruster 2018). The Rasborinae

encompass ca. 80 species in Sundaland (Sholihah et al. 2020), representing ca. 75% of Rasborinae diversity (Froese and Pauly 2020). Recent genetic reappraisals of Rasborinae species diversity in the area confirmed Sundaland's species identity and distribution ranges, along with the recognition of an abundant cryptic diversity (Dahruddin et al. 2017; Hubert et al. 2019; Sholihah et al. 2020). Most cryptic lineages identified have very narrow range distribution suggesting landscape fragmentation mostly contributed to generate this diversity (Sholihah et al. 2020).

Here, we explore the phylogenetic relationships of Sundaland's Rasborinae through mitochondrial genome skimming (Straub et al. 2012; Dodsworth 2015), with the aim to examine the potential impact of PCFs and palaeorivers on species proliferation of this species-rich group. We addressed the following questions: (1) Did palaeorivers serve as corridors of dispersal between islands during Pleistocene low stands? (2) Did palaeoriver watersheds prompt allopatric divergence across their boundaries? (3) Did PCF affect rates of species diversification? Through a dense taxonomic, spatial and mitochondrial genomic sampling within the group, the rasborine's tree of life was reconstructed and speciation events dated to explore dynamics of species proliferation through ancestral area estimations and model-based approaches.

Materials and Methods

Analytical procedure and sampling

The species diversity of the subfamily Rasborinae has been recently revisited in Sundaland through standardized DNA-based approaches (Dahruddin et al. 2017; Hubert et al. 2019; Sholihah et al. 2020). An ample, updated DNA barcode reference library has been made available by the authors (Sholihah et al. 2020). Here, the objective is to take advantage of this detailed genetic information to explore the Rasborinae tree of life. As such, the Rasborinae DNA barcode reference library of Sholihah et al. (2020) was used to guide taxon sampling for further mitogenome skimming (Straub et al. 2012; Dodsworth 2015). As the earliest branching events in the Rasborinae tree of life and major clades are still unknown (Brittan 1972; Liao et al. 2010), mitogenomes were first used to reconstruct a backbone phylogeny of the Rasborinae with the objective to identify major clades. Then, mitogenomes were combined to all DNA barcodes available to reconstruct detailed phylogenetic trees for

each major clade based on a dense taxon sampling. Sampling and collection management is as described in Sholihah et al. (2020). Specimens were captured using gears such as electrofishing, seine nets, cast nets and gill nets across sites that encompass the diversity of freshwater lentic and lotic habitats in Sundaland (Fig. 3.1). Specimens were identified following original descriptions where available, as well as monographs (Kottelat et al. 1993; Kottelat 2013) and further validated through DNA barcodes by including records from type localities (Hubert et al. 2019; Sholihah et al. 2020). Species names were further validated using several online catalogues (Eschmeyer et al. 2018; Froese and Pauly 2020). Specimens were photographed, individually labelled, and voucher specimens were preserved in a 5% formalin solution. Prior to fixation a fin clip or a muscle biopsy was taken and fixed separately in a 96% ethanol solution for further genetic analyses. Both tissues and voucher specimens were deposited in the national collections at the Museum Zoologicum Bogoriense (MZB), Research Center for Biology (RCB), and Indonesian Institute of Sciences (LIPI).

Sequencing

Genomic DNA was extracted using a MINIPREP SIGMA extraction kit following manufacturer's specifications. A 651-bp segment from the 5’ region of the cytochrome oxidase I gene (COI) was amplified as described in Sholihah et al. (2020). PCR amplifications were done using the primers cocktails C_FishF1t1/C_FishR1t1 including M13 tails (Ivanova et al. 2007) in a final volume of 10.0µl containing 5.0µl Buffer 2X, 3.3µl ultrapure water, 1.0µl each primer (10µM), 0.2µl enzyme Phire Hot Start II DNA polymerase (5U) and 0.5µl of DNA template (~50 ng). Amplifications were conducted as followed: initial denaturation at 98°C for 5 min followed by 30 cycles denaturation at 98°C for 5s, annealing at 56°C for 20s and extension at 72°C for 30s, followed by a final extension step at 72°C for 5 min. The PCR products were purified with ExoSap-IT (USB Corporation, Cleveland, OH, USA) and sequenced in both directions. The sequences and collateral information are available in BOLD (Ratnasingham and Hebert 2007) in the data set DS-BIFRA (Table S3.1, dx.doi.org/10.5883/DS-BIFRA).

Genomic libraries for mitogenome skimming were prepared following the protocol developed by Tilak et al. (2015) for multiplexed Illumina sequencing. Genomic DNA was physically fragmented through ultrasound (35 kHz) for a duration varying

between 10 and 20 min. We followed the Illumina library preparation procedure with blunt-end repair, adapter ligation, adapter fill-in and indexing PCR steps (13 cycles) developed by Meyer and Kircher (Meyer and Kircher 2010). Each step was followed by a purification using SPRI bead suspensions (Agencourt® AMpure® XP), adding 1.7 volume of Agencourt® AMpure® XP reagent per volume of sample and eluted in 25μl of ultra-pure water. Quantification of DNA libraries was done with a Nanodrop ND-800 spectrophotometer (Nanodrop technologies). Indexed libraries were pooled using their relative concentrations to ensure equimolarity and a single pool was single-read sequenced (150 bp long reads) on Illumina HiSeq 2500 at MGX (Montpellier, France). Mitogenomes were then assembled by reference to the closest mitogenome available among the 10 Rasborinae mitogenomes available in GenBank on Unipro UGENE (Okonechnikov et al. 2012). Complete mitogenomes were then annotated using the online tool MitoAnnotator (Iwasaki et al. 2013) available at mitofish.aori.u-tokyo.ac.jp. Annotated mitogenomes are accessible in GenBank (Table S3.1).

Reconstructing a backbone phylogeny of Rasborinae through mitogenomes

The Most Recent Common Ancestor (MRCA) of the subfamily Rasborinae has been previously estimated at 43 Ma (Betancur-R et al. 2017). Thus, tRNAs and Control Region were trimmed for phylogenetic reconstructions due to their fast substitution rates, and likely high levels of homoplasy. Protein and ribosomal RNA (rRNA) coding regions were retained and individually partitioned in subsequent phylogenetic reconstructions. First, a maximum likelihood (ML) tree was reconstructed using a partitioned model for each protein and rRNA coding regions with a GTR+I+Γ model as implemented in RAxML-HPC Blackbox (Miller et al. 2010) with RAxML 8 (Stamatakis 2014). Topological support was estimated with 5,000 non-parametric bootstraps replicates. Second, a calibrated tree was reconstructed using Bayesian inferences as implemented in BEAST 2.6.2 (Heled and Drummond 2010; Bouckaert et al. 2014). The most likely substitution models were jointly determined for all partitions using ModelFinder (Kalyaanamoorthy et al. 2017) as implemented in IQTREE online webserver (Nguyen et al. 2015) at http://iqtree.cibiv.univie.ac.at. The selected models were further used to conduct a Bayesian partitioned analysis based on a Yule model (uniform birth rate), relaxed clock with log normal distribution, and standardized site models as implemented in the SSM package in BEAST 2.6.2. Two Monte Carlo Markov

Chain (MCMC) of 50 million generations (burnin of 10%) were conducted to check for convergence and ESS estimates reached 200 using Tracer 1.7.1 (Drummond et al. 2012). Two clock rates were jointly estimated for rRNA and protein-coding regions along tree topology. MCMC were initiated with a 0.3% of divergence per million years (Myrs) for rRNA (Ortí and Meyer 1997; Hardman and Lundberg 2006) and 1.2% per Myrs for protein-coding regions (Bermingham et al. 1997). Independent runs were then combined using LogCombiner 2.6.2 (Bouckaert et al. 2014) and the maximum clade credibility tree, median age estimates and corresponding 95% highest posterior density (HPD) were summarized using TreeAnnotator 2.6.2 (Bouckaert et al. 2014). Both ML and Bayesian inferences were rooted using an assortment of mitogenomes (Table S3.1) available for several closely related subfamilies of Danionidae as well as other Cypriniformes families following previously published phylogenetic hypotheses (Tang et al. 2010; Betancur-R et al. 2017).

DNA barcodes, genetic species delimitation and species trees

Once major clades were identified within the subfamily, all DNA barcodes available from previous studies were compiled (Table S3.1). DNA barcode sequences were selected according to a preliminary screening of their phylogenetic affinities and robustness of the phylogenetic inferences. A Neighbour Joining (NJ) was first reconstructed for the 1,097 DNA barcodes from Sholihah et al. (2020) and branching support was estimated through 5,000 bootstrap replicates using PAUP 4.0a (Swofford 2001). Only DNA barcode records related to species with mitogenomes available, with bootstrap proportions (BP) above 80% were retained. Genetic delimitation of species follows the protocol described in Sholihah et al. (2020). Four different sequence-based methods of species delimitation were used to delimitate Molecular Operational Taxonomic Units (MOTUs) (Blaxter et al. 2005) using the 1,097 DNA barcodes. These methods were: (1) Refined Single Linkage (RESL) as implemented in BOLD and used to generate Barcode Index Numbers (BIN) (Ratnasingham and Hebert 2013), (2) Automatic Barcode Gap Discovery (ABGD) (Puillandre et al. 2012), (3) Poisson Tree Process (PTP) in its multiple rates version (mPTP) as implemented in the stand-alone software mptp_0.2.3 (Zhang et al. 2013), and (4) General Mixed Yule-Coalescent (GMYC) in its multiple rate version (mGMYC) as implemented in the R package Splits 1.0-19 (Fujisawa and Barraclough 2013). RESL and ABGD used DNA alignments as

input files, while a ML tree was used for mPTP and a Bayesian Chronogram based on a strict-clock model using a 1.2% of genetic distance per Myrs for mGMYC. The ML tree for mPTP was reconstructed using RAxML 8 using a GTR+Γ substitution model (Stamatakis 2014) and the ultrametric and fully resolved tree for mGMYC was reconstructed using BEAST 2.6.2 with two independent Markov chains of 50 million generations each including a Yule pure birth model tree prior, a strict-clock model and a GTR+I+Γ substitution model. Both runs were combined using LogCombiner 2.6.2 and the maximum clade credibility tree was constructed using TreeAnnotator 2.6.2. Duplicated haplotypes were pruned for further species delimitation analyses.

Once DNA barcodes were selected and MOTUs delimited, DNA barcode and mitogenome alignments were concatenated for each of the major clades identified. These concatenated alignments were further used to analyse phylogenetic relationships within clades using the Bayesian analysis implemented in the StarBEAST2 package (Ogilvie et al. 2017) from the BEAST 2.6.2 suite (Bouckaert et al. 2014). This approach implements a mixed-model including a coalescent component within species and a diversification component between species that allows accounting for variations of substitution rates within and between species (Ho and Larson 2006). StarBEAST2 jointly reconstruct gene trees and species trees, and as such requires the designation of species, which were determined using the consensus of our species delimitation analyses. A preliminary analysis conducted by partitioning each rRNA and protein-coding regions resulted in unstable MCMC and very low ESS. Analyses were further conducted using a single partition including rRNA and protein-coding regions, GTR+I+Γ substitution model, uncorrelated log-normal species tree model (UCLN), and MCMC of 60 million generations. Age of each clade MRCA estimated from the initial backbone phylogeny was used as a calibration point with a normal distribution and a sigma of 1.0. Clock rate was estimated and an initial value of 0.8 % of divergence per Myrs was used according to the initial BEAST 2.6.2 analysis of the mitogenome dataset. Independent runs were combined using LogCombiner 2.6.2 (Bouckaert et al. 2014). Gene and species maximum clade credibility trees, median/mean age estimates and corresponding 95% HPD were summarized using TreeAnnotator 2.6.2 (Bouckaert et al. 2014).

Diversification rates estimation

Lineages through time (LTT) were plotted using the species trees of each major clades with the R-package *ape* (Paradis and Schliep 2019). Confidence intervals were computed using 1,000 dated trees sampled along the StarBEAST2 MCMC. To test the impact of past environmental dynamics on diversification, we relied on a ML framework with five diversification models (constant-rate, time-dependent, temperature-dependent, sea-level-dependent, and diversity-dependent models) and their variants (Condamine et al. 2013a, 2019). In total, we fitted 17 diversification models (Table S3.2) using the R-packages *RPANDA* 1.3 (Morlon et al. 2016) and *DDD* 3.7 (Etienne et al. 2012). We accounted for potential missing lineages in the phylogeny in the form of global sampling fraction, i.e. the ratio of sampled lineages diversity over the total described lineages, and ran these analyses for sampling fractions of 100% and 90%. In *RPANDA*, we first fitted two constant-rate models as initial references, namely: BCST (speciation rate constant through time with no extinction) and BCSTDCST (speciation and extinction rates constant through time). Second, we fitted four time-dependent models: BtimeVar (speciation rate varying through time with no extinction), BtimeVarDCST (speciation rate varying through time with constant extinction), BCSTDtimeVar (constant speciation and extinction rate varying through time) and BtimeVarDtimeVar (both speciation and extinction rates varying through time). Lastly, we fitted eight models with speciation and extinction rates varying according to external environmental variables: BtemperatureVar and Bsea-levelVar (speciation varying in function of the environmental variable), BtemperatureVarDCST and Bsea-levelVarDCST (speciation varying in function of the variable with constant extinction rate), BCSTDtemperatureVar and BCSTDsea-levelVar (extinction rate varying in function of the variable and constant speciation rate) and BtemperatureVarDtemperatureVar and Bsea-levelVarDsea-levelVar (both speciation and extinction rates varying in function of the variable).

Exponential dependence with time, temperature or sea level and diversification rates are chosen for its robustness and flexibility depending on the strength and direction of the dependence to the fitted variable. Speciation (λ) and extinction (μ) rates are parameterized as follows. When λ and μ are exponential functions of sea level (S) through time (t), the equations are $\lambda(S_{(t)}) = \lambda_0 \times e^{\alpha S(t)}$ and $\mu(S_{(t)}) = \mu_0 \times e^{\beta S(t)}$, where λ_0 and μ_0 are respectively the expected λ and μ at S = 0 meter, while α and β

are coefficients that measure the strength and the sign of the relationship with sea level (e.g. $\alpha > 0$ and $\beta > 0$ respectively indicate λ and μ increase with sea level high stands). Similar parameterisation can be used for exponential relationship between temperature (*T*) through time (*t*) and λ as well as μ rates in which $\lambda(T_{(t)}) = \lambda_0 + \alpha T_{(t)}$ or $\lambda(T_{(t)}) = \lambda_0 \times e^{\alpha T(t)}$, where *T(t)* is the temperature at time *t* and λ_0 is speciation rate at *T* = 0°C.

The last three diversity-dependent ML models were fitted using DDD package in which λ and μ vary as linear functions of number of lineages within each clade (Etienne et al. 2012). The diversity-dependent models are parameterised by λ_0, μ_0 (respectively indicating λ and μ at the absence of competing lineage), K (carrying capacity, representing asymptotic clade size). All λ and μ were constrained to be positive. The results were then compiled with the previous 14 models in RPANDA and all are compared using corrected Akaike Information Criterion (AICc) and Akaike weights (AICω). The model with the lowest AICc and highest AICω was considered as the best fitting model for the phylogeny.

Ancestral areas estimation

To explore dispersal and vicariance history based on the Palaeoriver Hypothesis, we reconstructed ancestral distribution of Sundaland Rasborinae using R-package *BioGeoBEARS* 1.1.2 (Matzke 2013, 2014) based on the StarBEAST2 species trees. Species presence/absence were compiled (Table S3.1) for two sets of geographical delimitation based on: (1) palaeoriver and (2) contemporary island boundaries. Then, geographical patterns of speciation were recorded as follows: (1) no dispersal, sister species co-occur within the same palaeoriver and the same island; (2) dispersal between islands within a palaeoriver, sister species are alternatively distributed on different islands within the same palaeoriver; (3) dispersal between palaeorivers within the same island, sister species are alternatively distributed on different palaeorivers within the same islands; and (4) dispersal between islands and between palaeorivers, sister species are alternatively distributed on different palaeorivers and different islands. Ancestral area estimations involving palaeorivers were based on the following geographical areas (Fig. 3.1): (1) the Malacca Straits palaeoriver, (2) the East Sunda palaeoriver, (3) the North Sunda palaeoriver, (4) the Siam palaeoriver, (5) the Northern Borneo river system, and (6) the Mekong river

system. On the other hand, estimations based on insular delimitation followed geographical divisions of: (1) Sumatra-Bangka-Belitung, (2) Java-Bali-Lombok, (3) Borneo, and (4) Mainland Southeast Asia. For analytical requirements, only MOTUs with known locality(ies) that could be associated to both geographical frameworks were used.

Inferences of ancestral areas using BioGeoBEARS were conducted using six alternative models including dispersal-extinction cladogenesis (DEC), DEC+J, dispersal-vicariance analysis ML-version (DIVALIKE), DIVALIKE+J, Bayesian biogeographical inference model (BAYAREALIKE) and BAYAREALIKE+J (Matzke 2014; van Dam and Matzke 2016). The inclusion of the parameter J has been recently criticized from a conceptual and statistical perspective (Ree and Sanmartín 2018). The concept of jumping dispersal has been developed for insular systems to account for the settlement of a new lineage established by colonization without an intermediate widespread ancestor (Clark et al. 2008; Ree and Sanmartín 2018). Considering biogeographical scenario of Sundaland and insularity of the system, jumping dispersal cannot be discarded *a priori* from a conceptual perspective and several studies have previously pinpoint the importance of jumping dispersal in insular systems (Cowie and Holland 2006; de Bruyn et al. 2013; Condamine et al. 2015b; Beck et al. 2017; Hendriks et al. 2019). Models of ancestral area estimation including the J parameters were thus considered here and the best-fit model was estimated using the AICc.

Results

Phylogenetic reconstructions

A total of 58 new mitogenomes were successfully assembled for four genera including three *Brevibora*, one *Pectenocypris*, 45 *Rasbora* and seven *Trigonopoma*. Mitogenomes were ca. 16,500 bp long on average and include the coding regions of two rRNA (12S, 16S), 22 tRNA, 13 protein-coding genes and the Control Region (CR). In addition, 10 Rasborinae mitogenomes were retrieved from GenBank including the genera *Amblypharyngodon*, *Horadandia*, *Rasboroides*, *Rasbora* and *Boraras* (Table S3.1). These 10 mitogenomes were used as reference genomes for assembly and further phylogenetic reconstructions. An additional set of 11 mitogenomes was retrieved from GenBank for additional Danionidae subfamilies and Cypriniformes families, which were used as outgroup (Table S3.1). For phylogenetic reconstruction,

tRNA coding regions and CR were trimmed. The final alignment included 79 mitogenomes and 13,898 bp consisting of 2,518 bp of rRNA and the 13 protein-coding regions including ND1 (975 bp), ND2 (1,045 bp), COI (1551 bp), COII (691 bp), ATP8 (1655 bp), ATP6 (673 bp), COIII (785 bp), ND3 (349 bp), ND4L (297 bp), ND4 (1,375 bp), ND5 (1,822 bp), ND6 (518 bp), and Cytb (1,134 bp).

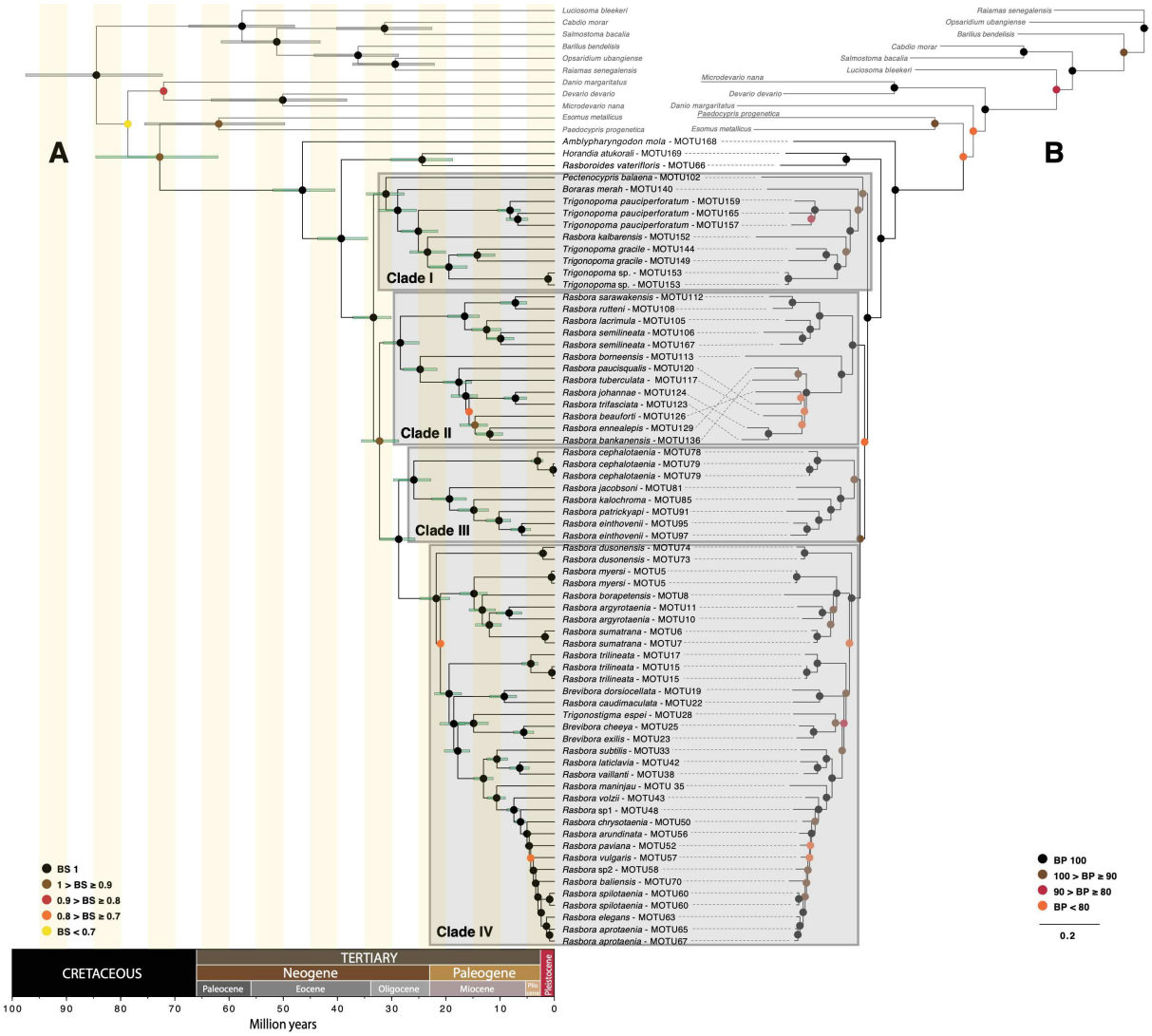

Figure 3.2 Phylogenetic reconstructions in Rasborinae based on 79 mitogenomes. A, Bayesian Maximum Clade Credibility Tree based on 14 partitions (13 protein coding and one rRNA coding partitions), two clock rates (0.3% per Myrs for rRNA partition, 1.2% per Myrs for the 13 protein coding partitions) and including posterior probabilities (PP) and confidence intervals of age estimates. B, Maximum Likelihood phylogenetic tree based on 14 partitions (13 protein coding and one rRNA coding partitions) and including bootstrap proportion (BP).

Phylogenetic reconstructions based on the 79 mitogenomes are generally well supported with most internal branching events supported by posterior probabilities (PP) of 100% for the Bayesian inference (Fig. 3.2a) and Bootstrap Proportions (BP) above 90% for the ML reconstruction (Fig. 3.2b). Bayesian and ML topologies are

highly congruent with continental Asian Rasborinae (Amblypharingodon, Horadandia and Rasboroides) corresponding to the earliest branching events in Rasborinae tree and Sundaland lineages constituting a monophyletic group (Fig. 3.2). Four major clades are identified within Sundaland: (1) Clade I including the genera Pectenocypris, Boraras, Trigonopoma and Rasbora kalbarensis; (2) Clade II including some Rasbora species; (3) Clade III including some Rasbora species; and (4) Clade IV including some Rasbora species and the genera Brevibora and Trigonostigma. The MRCA of the subfamily Rasborinae is traced back to 46.50 Ma (95% HPD = 40.53-51.93 Ma, Eocene), while the MRCA of the four Sundaland clade is dated at 33.41 Ma (95% HPD = 30.15-37.19 Ma, Oligocene-Eocene transition). The four clades have varying age estimates with the MRCA of Clade I dated at 31.10 Ma, (95% HPD = 27.76-34.725 Ma, Oligocene), the MRCA of Clade II at 28.44 Ma (95% HPD = 24.98-31.61 Ma, Oligocene), the MRCA of Clade III at 25.95 Ma (95% HPD = 22.81-29.68 Ma, Oligocene) and the MRCA of Clade IV at 21.83 Ma (95% HPD = 19.36-24.86 Ma, early Miocene). Estimated clock rates ranged between 0.31% per Myrs, with a variance of 0.0017%, for rRNA coding regions and 0.74% per Myrs, with a variance of 0.0048%, for protein-coding regions.

The four clades have a substantial proportion of cryptic diversity (Fig. 3.3) as previously reported (Sholihah et al. 2020). DNA barcodes selection based on 80% BP threshold yielded 1,017 DNA barcode sequences, which together with the 66 Rasborinae mitogenomes sum up to 1,083 sequences for 71 nominal species, 157 MOTUs and 10 genera of Rasborinae (Table S3.1). ML and Bayesian gene trees are congruent for all clades (Fig. 3.3). Most internal branching events are well supported in the ML gene trees (Figs. 3.3a, 3.3c, 3.3e, 3.3g), except among some *Trigonopoma* MOTUs within Clade I (Fig. 3.3a), and the most derived MOTUs of Clade IV (Fig. 3.3g). Generally, Bayesian gene trees were more supported for all clades (Figs. 3.3b, 3.3d, 3.3f, 3.3h) with most internal branching events supported by PP>0.9. The estimated mitogenome clock rate was 0.55% per Myrs for Clade I, 0.64% per Myrs for Clade II, 0.51% per Myrs for Clade III, and 0.56% per Myrs for Clade IV. MRCAs age estimates were very similar compared to the backbone phylogeny with the MRCA of Clade I dated at 31.08 Ma (95% HPD = 29.26-32.97 Ma), MRCA of Clade II at 27.99 Ma (95% HPD = 26.32-29.72 Ma), MRCA of Clade III at 24.50 Ma (95% HPD = 22.63-26.37 Ma) and MRCA of Clade IV at 21.89 Ma (95% HPD = 20.49-23.46 Ma). Individual clade phylogenetic reconstructions confirm the monophyly of the genera *Boraras* (Figs. 3.3a

& 3.3b), *Pectenocypris* (Figs. 3.3a & 3.3b) and *Trigonostigma* (Figs. 3.3g & 3.3h), the paraphyly of the genus *Trigonopoma* (Figs. 3.3a & 3.3b), and the polyphyly of the genera *Brevibora* (Figs. 3.3g & 3.3h) and *Rasbora* (Fig. 3.3).

Temporal and spatial diversification trends

MOTUs age estimates highlight that nearly half have Pleistocene origins (47.7%) by contrast with only 10.3% of nominal species (Fig. 3.4, Fig. S3.1). However, this proportion varies among clade with 55% of MOTUs with Pleistocene origin in Clade IV, 50% in Clade III, 41% in Clade I and 36% in Clade II. Diversification models indicate no clade-specific patterns of diversification (Table S3.3). The constant speciation rate without extinction (BCST) model is the most likely for all clades (AICω of 0.23 for Clades I and II, 0.262 for Clade III, and 0.22 for Clade IV), with both sampling fractions used (Table 3.1, Table S3.3). Speciation rates are similar for Clades I, II and III with λ = 0.0898, 0.108, 0.1195 events/Myr/lineage, respectively (Table 3.1, Fig. S3.1). Clade IV, however, shows a higher rate of speciation with λ = 0.147 (Table 3.1, Fig. S3.1).

A total of 139 MOTUs from 60 nominal species with known geographical distribution were used for ancestral area estimations for the four clades (Fig. 3.4). The most likely biogeographical models include the J parameter for all clades and both island-based and palaeoriver-based geographical partitioning (Table S3.4). DEC+J is the most likely model in most cases, except for Clade III with the palaeoriver-based partitioning supporting the DIVALIKE+J model. Ancestral area estimations point to Borneo as the most likely origin of Clades II and IV and subclades of Clade II, with a high probability of Bornean ancestry of their MRCAs (Figs. 3.4f, 3.4g, 3.4h). The insular ancestry of Clade I is not resolved, but the North Sundaland palaeoriver is identified as the most likely origin of Clade I (Fig. 3.4e) as well as Clades II and IV (Figs. 3.4f, 3.4h). Ancestral areas estimation indicates a recent colonization of Java during the Pliocene (3.807 Ma) by Clade IV (Fig. 3.4h). *In situ* diversification is observed in all clades, particularly in Borneo (island-based analyses) and North Sunda (in palaeoriver-based analyses) as exemplified by the upper group of Clade II (*Rasbora lacrimula, R. hubsii, R. semilineata, R. sarawakensis, R. rutteni*) from Borneo (15.80 Ma, 95% HPD = 14.22-17.53 Ma), as well as *Trigonopoma gracile* MOTU144-MOTU148 (5.99 Ma, 95% HPD = 3.08-9.495 Ma) and *Trigonopoma pauciperforatum* MOTU154-MOTU157 (3.54 Ma, 95% HPD = 1.71-5.94 Ma) from North Sunda (Fig. 3.4).

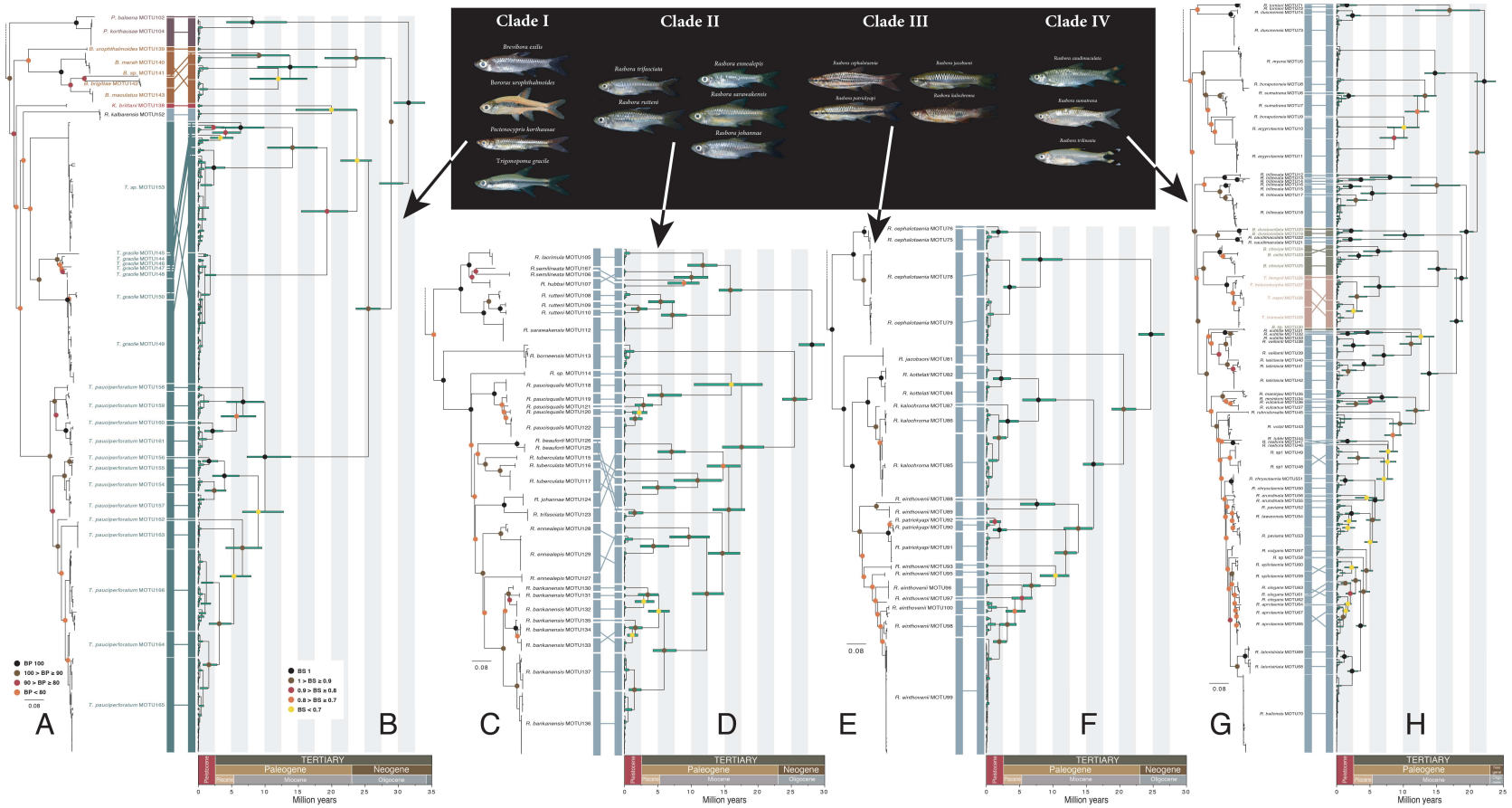

Figure 3.3 Mitochondrial gene trees of Clades I, II, III and IV.
Maximum Likelihood trees and bootstrap proportions (BP) are in panels A, C, E, and G, or Clades I, II, III and IV, respectively. Bayesian Maximum Clade Credibility Trees and posterior probabilities (PP) are in panels B, D, F, and H for Clades I, II, III, and IV, respectively. All ML trees were rooted using A. mola, R. vaterifloris and H. atukorali, Clade I ML tree (A) included additional extra-groups as follows: R. cephalotaenia, R. einthovenii, R. sumatrana, R. dusonensis, R. aprotaenia, R. semilineata, R. rutteni, R. bankanensis, and R. borneensis. Clade II ML tree (C) included additional extra-groups as follows: R. cephalotaenia, R. einthovenii, R. sumatrana, R. dusonensis, R. aprotaenia, B. maculatus, T. n sp, and P. balaena. Clade III ML tree (E) included additional extra-groups as follows: B. maculatus, T. n sp, P. balaena, R. borneensis, R. bankanensis, R. semilineata, R. rutteni, R. sumatrana, R. dusonensis, and R. aprotaenia. Clade IV ML tree (G) included additional extra-groups as follows: B. maculatus, T. n sp, P. balaena, R. borneensis, R. bankanensis, R. semilineata, R. rutteni, R. cephalotaenia, and R. einthovenii.

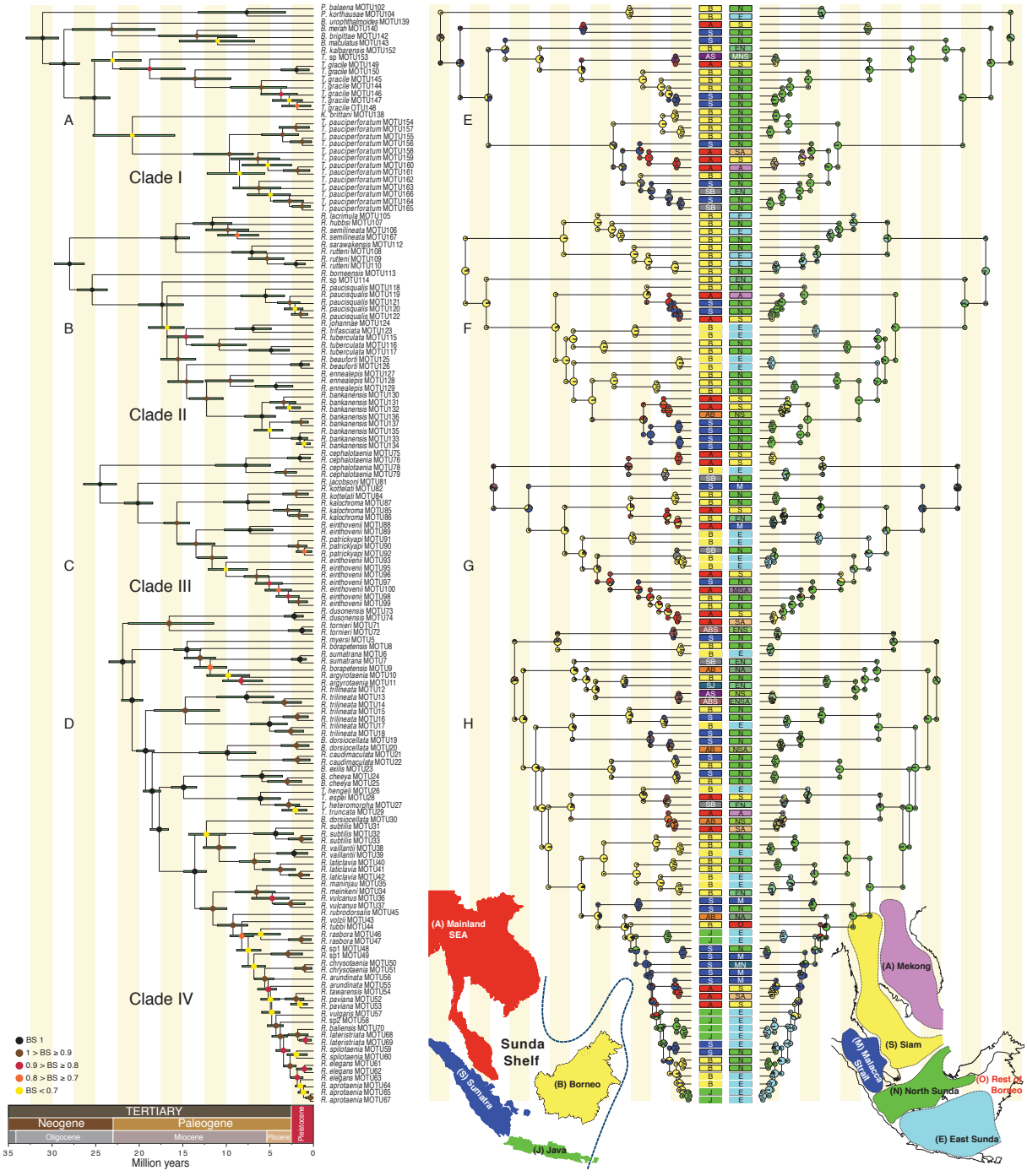

Figure 3.4 Mitochondrial species trees and ancestral area estimations of Clades I, II, III, and IV according to an island-based or a palaeoriver-based geographical partitioning.
A to D, species trees for Clades I, II, III and IV, respectively. E to H, ancestral area estimations for Clades I, II, III, and IV, respectively.

Table 3.1 Summary statistics of the most likely diversification models for Clades I, II, III, and IV including acronym, rate variation, number of parameters (NP), speciation rates (λ), corrected Akaike Information Criterion (AICc) and Akaike weight (AICω)

Clade	Age (Myr)	Extant Diversity (#MOTU)	Assumed Sampling Fraction (%)	Most Likely Model	Rate variation	NP	λ	logL	AICc	Akaike ω
Clade I	31.08	29	100	BCST	constant	1	0.0851	-93.529	189.206	0.23
Clade I	31.08	29	90	BCST	constant	1	0.0898	-93.586	189.32	0.221
Clade II	27.99	33	100	BCST	constant	1	0.102	-101.758	205.644	0.23
Clade II	27.99	33	90	BCST	constant	1	0.108	-101.689	205.506	0.245
Clade III	24.5	22	100	BCST	constant	1	0.113	-63.608	129.416	0.262
Clade III	24.5	22	90	BCST	constant	1	0.1195	-63.666	129.532	0.255
Clade IV	21.89	69	100	BCST	constant	1	0.1466	-195.656	393.371	0.221
Clade IV	21.89	69	90	BCST	constant	1	0.1548	-195.559	393.178	0.238

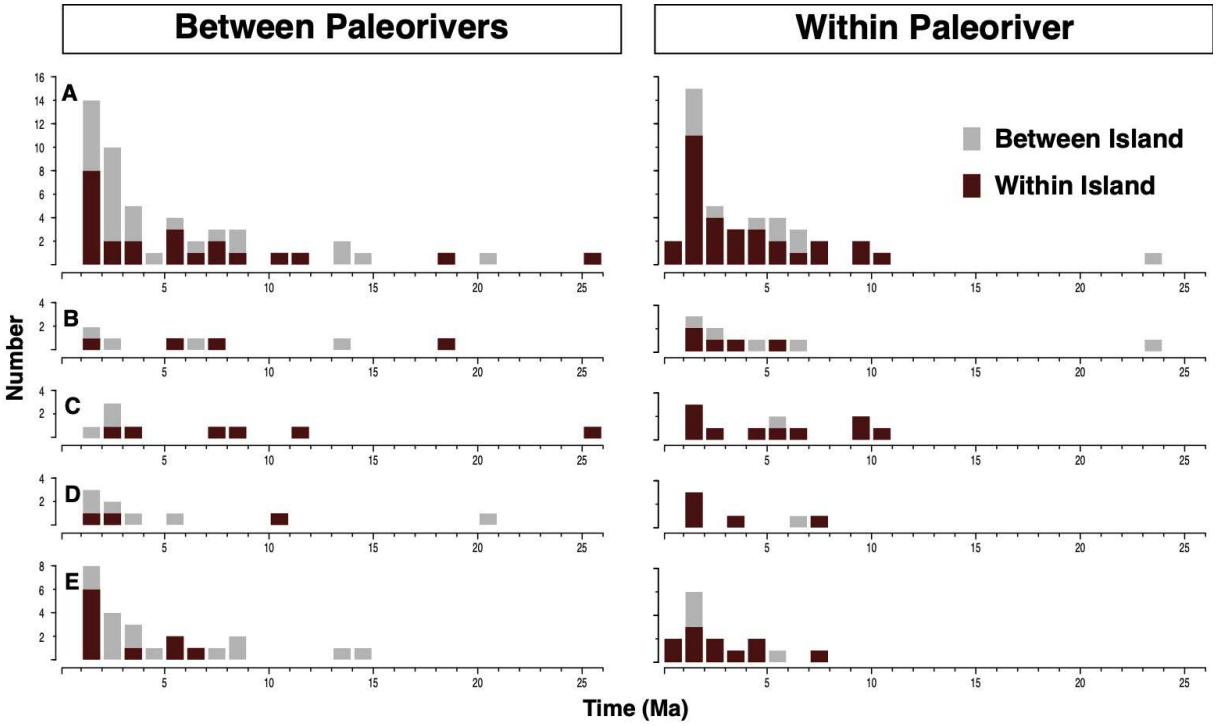

Figure 3.5 Distribution of speciation events through time according to geographical patterns for all clades (A), Clade I (B), Clade II (C), Clade III (D), and Clade IV (E)

In total, most speciation events are observed within islands with 59% and between palaeorivers with 54% (Table 3.2, Fig. 3.5). Most speciation events are associated to dispersal as 66% involve either different islands or different palaeorivers, while 34% occurred within the same island and the same palaeoriver (Table 3.2). However, these trends vary through time. Most speciation between islands occurs

during the last 5 Myrs and is mostly associated to dispersal between palaeorivers (Fig. 3.5a). Along the same line, *in situ* speciation within palaeoriver occurs mostly within island during the last 5 Myrs (Fig. 3.5a). Geographical patterns of speciation vary among clades with a spectacular dominance of speciation within islands for Clade II, most frequently occurring within the same palaeoriver (Table 3.2, Fig. 3.5c). Proportions are more balanced for other clades (Table 3.2), with most speciation events within palaeorivers occurring within islands (Fig. 3.5).

Table 3.2 Summary statistics of geographical patterns of speciation events for Clades I, II, III, and IV

Taxa	Islands		Total (%)
	within (%)	between (%)	
All	59	41	-
within palaeorivers	34	12	46
between palaeorivers	25	29	54
Clade I	50	50	-
within palaeorivers	28	28	56
between palaeorivers	22	22	44
Clade II	80	20	-
within palaeorivers	50	5	55
between palaeorivers	30	15	45
Clade III	53	47	-
within palaeorivers	33	7	40
between palaeorivers	20	40	60
Clade IV	54	46	-
within palaeorivers	28	10	38
between palaeorivers	26	36	62

Discussion

Sundaland biogeography

Our phylogenetic and biogeographic reconstructions are in line with geological reconstructions in Southeast Asia (Hall 2009; Hall et al. 2011; Lohman et al. 2011). Rasborinae lineages from Mainland Asia are associated with the earliest branching events in the Rasborinae tree. Our molecular dating of Sundaland clades indicates that the four Sundaland lineages started to diversify between 31.08 and 21.89 Ma, a timeframe matching the onset of Borneo isolation from mainland Asia (Fig. 3.1). Ancestral area estimations further confirm this temporal match as Borneo is inferred

as the most likely origin of most Sundaland clades (Fig. 3.4). These results suggest Sundaland Rasborinae originate from Mainland Asia and colonized Sundaland during the earliest stages of its geomorphological history in the Oligocene. This timeframe supports the pre-Pleistocene colonisation of Sundaland by freshwater fishes (Dodson et al. 1995; de Bruyn et al. 2013; Hendriks et al. 2019; Sholihah et al. in press).

Our biogeographic estimations further highlight the importance of the North Sunda palaeoriver during the initial freshwater fish diversification in Sundaland (de Bruyn et al. 2013, 2014; Sholihah et al. in press). The North Sunda palaeoriver is the most likely centre of origin of most Rasborinae clades (Fig. 3.4). The diversification of Rasborinae further followed Sundaland's geological history with varying colonization scenarios for Sumatra and Java. Java is the youngest of Sundaland islands, with a separation that occurred during the last 5 Myrs (Fig. 3.1), which is consistent with our molecular dating as none of the MOTUs endemic of Java are older than 4.06 Myrs (*Rasbora* sp2). However, our reconstructions indicate that colonization of Java likely results from three distinct immigration events (Fig. 3.4h) followed by *in situ* diversification. The occurrence of immigration followed by *in situ* speciation has been previously suggested for freshwater fishes in Java, including *Rasbora* (Kusuma et al. 2016; Hutama et al. 2017; Hubert et al. 2019). These diversification events are likely resulting from the intense volcanic activity in Java, which formed several volcanic arches and fragmented rivers into multiple, small and confined watersheds. Java is the only Sundaland island with a predominant influence of volcanic activity during its emergence (Lohman et al. 2011). These rugged aquatic landscapes likely fragmented ancestral lineages, creating local radiations within the island (Nguyen et al. 2008; Pouyaud et al. 2009; Hubert et al. 2015, 2019; Kusuma et al. 2016; Dahruddin et al. 2017; Hutama et al. 2017; Sholihah et al. in press).

This scenario in Java contrasts with Sumatra, which biogeographic estimations suggest a more ancient and intricate scenario of colonization (Fig. 3.4). Two ancient lineages are detected in Sumatra: *Rasbora kalbarensis* diverged 23.05 Ma, and *R. jacobsoni* diverged 20.15 Ma. These lineages occur in the North Sunda or Malacca palaeorivers (Fig. 3.4). The age estimates of these species contrast with most lineages in Sumatra, not exceeding 13.38 Myrs and mostly tracing back to the Miocene-Pliocene transition ca. 5 Ma. These contrasted patterns suggest two waves of colonization during: (1) the onset of Borneo isolation and development of the North Sunda palaeoriver, and (2) the final stage of isolation of Sumatra, when land bridges

were still connecting the Southern tip of Sumatra, Borneo and West Java (Fig. 3.1d). Cases of *in situ* diversification are also detected in Sumatra (Fig. 3.4).

Dispersal and Pleistocene palaeoenvironments

Our ancestral area estimations and modelling of vicariance and dispersal indicate jump dispersal is common in Rasborinae, a result supported by the significant increase of likelihood scores and Akaike weights when the J parameter is included (Table S3.4). Multiple cases of trans-island dispersal have been reported among Sundaland freshwater fishes (Pouyaud et al. 2009; Adamson et al. 2010; de Bruyn et al. 2013; Tan and Lim 2013; Lim et al. 2016b; Beck et al. 2017; Nurul Farhana et al. 2018). However, our reconstructions question Pleistocene eustatic fluctuations as main driver of dispersal between islands. During glacial maxima, sea levels dropped (Miller et al. 2005). The shallow Java Sea has given way to exposed lands, which likely had freshwater drainage systems that extended between modern islands (Voris 2000; Woodruff 2010). These palaeoriver systems are expected to ease dispersal between islands and promote speciation. This prediction received little support here. Most speciation events occurred within islands (Table 3.2) and most speciation events between islands occurred between palaeorivers (Fig. 3.5). This trend of speciation within islands is particularly marked for Clade II, in which 80% of speciation events occurred within islands (Table 3.2). The predominance of speciation within islands is linked to Rasborinae ecology. Most species in Clades I, II, and III are small-size species inhabiting forested streams and peat swamps (Kottelat et al. 1993; Sholihah et al. 2020). These species are forest-dependent taxa.

During glacial maxima, climate was cooler and drier. Savanna and seasonal forest corridors expanded through the interior of Sundaland (Heaney 1991; Bird et al. 2005), enhancing microclimates and diversity of freshwater habitats along the inter-island channels (Heaney 1991; Gorog et al. 2004; Bird et al. 2005; Pouyaud et al. 2009; Wurster et al. 2019). While documented for terrestrial organisms, speciation pattern of Clade II suggests that vegetational changes during glacial maxima also limited dispersal for aquatic, forest-associated organisms. Clade II further exemplifies the intricate interactions between palaeoenvironments and dispersal. Habitat specificity toward forest habitats and peat swamps may have limited dispersal between islands despite the availability of freshwater corridors. These results highlight the

dynamic interactions between palaeoenvironments and Pleistocene eustatic fluctuations (Esselstyn and Brown 2009; Brown et al. 2013; Papadopoulou and Knowles 2015a, 2015b; Sholihah et al. in press).

Macroevolutionary drivers of diversification

Pleistocene climatic fluctuations have been frequently invoked to account for Pleistocene increased rates of diversification (Mittelbach et al. 2007; Weir and Schluter 2007). In Sundaland, Pleistocene eustasy is predicted to induce cycle of dispersal during glacial time and vicariance during interglacial periods (Kottelat et al. 1993; Voris 2000; Woodruff 2010). An amplification of sea-level fluctuations during the Pleistocene has been observed (Fig. S3.1) (Miller et al. 2005), which predict increased speciation rate in the context of Sundaland. This prediction is not supported here. Constant speciation models of diversification are the most likely for all clades (Table S3.3). In a model with a constant probability of speciation per lineage through time, species diversity increases linearly over time without levelling off toward the present (Table S3.2). Thus, the Pleistocene origin of a large proportion of MOTUs could be a consequence of constant diversification through time. This result might be surprising when considering the vast array of aquatic habitats occupied by the Rasborinae in Sundaland: lowland vs. highland, fast vs. slow running waters, peat swamps, lakes, ponds (Brittan 1972; Kottelat et al. 1993; Kusuma et al. 2016; Hubert et al. 2019; Sholihah et al. 2020). Besides, ecological transitions are scarce in Rasborinae. Clades I, II and III include small-size, forest-associated species while Clade IV include all the large, riverine and open-habitat *Rasbora* species (Kottelat et al. 1993; Liao et al. 2011; Sholihah et al. 2020). This pattern suggests that adaptive habitat-shift had limited impact on the pace of diversification in Rasborinae. Yet, Clade IV has the highest speciation rate of the four clades and is also the only clade that colonized Java. Altogether, the transition between forest and open habitat that happened during the onset of Clade IV diversification and higher speciation rate suggest an influence of ecological contingency on Rasborinae diversification. This group of riverine and opportunistic species was successful at colonizing open habitats, a previously uncolonized set of habitats for Rasborinae.

Despite the high diversity of MOTUs in Sundaland Rasborinae, we find no evidence of diversity equilibrium. The diversity-dependent diversification models

received little support for all clades, suggesting intra-clade biotic interactions have little influence on either speciation opportunities or species' probabilities of maintenance (Alonso et al. 2006; Hubert et al. 2015a). If considering the ample distribution and abundance of Rasborinae in Sundaland and their presence in all aquatic habitats, this result is surprising. Rasborinae assemblages were likely unsaturated during the diversification of the subfamily (Cornell 1993). Diversity equilibrium and slowdown of diversification rates implies a shift in speciation/extinction equilibriums once assemblages are saturated (Phillimore and Price 2008; Kisel et al. 2011; Hubert et al. 2015a). Such dynamics are underlined by the establishment of speciation/extinction equilibriums. Here, PCFs might be expected to have cyclically perturbed aquatic assemblages, and regularly disrupted a potential course toward equilibrium. Alternatively, Sundaland ecological carrying capacity might still be far from being reached, despite the staggering species richness of Sundaland's ichthyofauna.

Robustness of the inferences and systematic implications

Our study first confirms the monophyly of the subfamily Rasborinae and its distinctiveness from the subfamilies Chedrinae, Danioninae and Esominae (Mayden et al. 2007; Rüber et al. 2007; Conway et al. 2008; Fang et al. 2009; Tang et al. 2010; Tan and Armbruster 2018). The estimated age of the Rasborinae MRCA, dated at 46.50 Ma (95% HPD = 40.53-51.93 Ma), is consistent with the 43.58 Myrs reported by Betancur-R et al. (2017) based on one mitochondrial and 20 nuclear genes. Our phylogenetic reconstructions reveal that *Rasbora* species are scattered across all clades, a result in line with previous molecular and morphology-based phylogenetic studies, which highlighted that *Rasbora* is encompassing lineages, morphologically very similar, but of distinct evolutionary origins (Liao et al. 2010, 2011; Lumbantobing 2010; Tang et al. 2010; Tan and Armbruster 2018; Sholihah et al. 2020). Yet, most *Rasbora* species groups initially described by Brittan (Brittan 1954, 1972) are recovered. For instance, Clade III matches the delimitation of *R. einthovenii* species group and Clade II matches the boundaries of *R. trifasciata* species group (Brittan 1972; Liao et al. 2010). The elevation of *R. pauciperforatum* species group to the genus level (*Trigonopoma*) by Liao et al. (2010) is supported here. However, our results indicate *R. kalbarensis* should probably be considered a member of the genus *Trigonopoma*. Likewise, the genus *Boraras* described by Kottelat and Vidthayanon

(Kottelat and Vidthayanon 1993) is supported here. Most of the taxonomic conflicts are concentrated in Clade IV, in which *R. lateristriata* and *R. sumatrana* species groups from Brittan (1972), and the genus *Brevibora* from Liao et al. (2011) are not monophyletic. The present study warrants further taxonomic works within Rasborinae and highlights the need of an in-depth revision of the genera *Rasbora* and *Brevibora*.

Conclusion

Our study shows that PCF and Pleistocene eustasy had less impact on rasborine diversification than expected under the Palaeoriver Hypothesis. Geographical patterns of speciation suggest that limited dispersal abilities and *in situ* diversification were predominant during the diversity build-up of the Rasborinae in Sundaland. Our phylogenetic and biogeographic reconstructions are in line with the timeframe of geological reconstructions in Southeast Asia. The Rasborinae followed the geological history of Sundaland and tightly matched the onset of isolation of Borneo. In particular, the North Sunda palaeoriver is identified as a key aquatic system during the rise of the subfamily and an important source of diversity for the neighbouring river systems. Our study also provides new lines of evidence about dispersal of freshwater organisms in Sundaland. PCFs poorly explain geographical patterns of speciation and dispersal between islands. Surprisingly, our macroevolutionary inferences show no evidence of diversification slowdown and diversity ceiling, despite the exceptional levels of species richness of Sundaland ichthyofauna. This unexpected result further questions mechanisms underlying these diversification trends. Several alternative scenarios may be invoked, including the impact of PCFs on disturbing speciation/extinction equilibriums. Our study warrants further research on the evolutionary mechanisms underlying diversification in such species-rich tropical systems.

Acknowledgements

The authors wish to thank Siti Nuramaliati Prijono, Bambang Sunarko, Witjaksono, Mohammad Irham, Marlina Adriyani, Ruliyana Susanti, Rosichon Ubaidillah, the late Renny K. Hadiaty, Hari Sutrisno and Cahyo Rahmadi at Research Centre for Biology (RCB-LIPI) in Indonesia; Edmond Dounias, Jean-Paul Toutain, Robert Arfi and Valérie Verdier from the ‘Institut de Recherche pour le

Développement'; Joel Le Bail and Nicolas Gascoin at the French embassy in Jakarta for their continuous support. We also would like to thank Eleanor Adamson, Hendry Budianto, Tob Chann Aun, Pak Epang, Herman Ganatpathy, Sébastien Lavoué, Michael Lo, Hendry Michael, Joshua Siow, Heok Hui Tan, Elango Velautham, Norsham S. Yaakob, and Denis Yong for their help in the field. We are also particularly thankful to Sumanta at IRD Jakarta for his help during the field sampling in Indonesia. The present study and all associated methods were carried out in accordance with relevant guidelines and regulation of the Indonesian Ministry of Research and Technology (Indonesia), the Economic Planning Unit, Prime Minister's Department (Malaysia), the Forest Department Sarawak (Malaysia), the Vietnam National Museum of Nature (Vietnam) and the Inland Fisheries Research and Development Institute (Cambodia). Field sampling in Indonesia was conducted according to the research permits 7/TKPIPA/FRP/SM/VII/2012, 68/EXT/SIP/FRP/SM/VIII/2013, 361/SIP/FRP/E5/Dit.KI/IX/2015, 50/EXT/SIP/FRP/E5/Dit.KI/IX/2016, 45/EXT/SIP/FRP/E5/Dit.KI/VIII/2017, and 392/SIP/FRP/E5/Dit.KI/XI/2018 for Nicolas Hubert, and, 1/TKPIPA/FRP/ SM/I/2011 and 3/TKPIPA/FRP/SM/III/2012 for Lukas Rüber. The Fieldwork in Peninsular Malaysia and Sarawak was conducted under permits issued by the Economic Planning Unit, Prime Minister's Department, Malaysia (UPE 40/200/19/2417 and UPE 40/200/19/2534) and the Forest Department Sarawak (NCCD.970.4.4[V]-43) and were obtained with the help of Norsham S. Yaakob (Forest Research Institute Malaysia, Kepong, Kuala Lumpur, Malaysia). Luong Van Hao and Pham Van Luc (Vietnam National Museum of Nature) helped with arranging research permits in Vietnam and So Nam (Inland Fisheries Research and Development Institute, IFReDI) helped with arranging research permits in Cambodia. All experimental protocols were approved by the Indonesian Ministry of Research and Technology (Indonesia), the Indonesian Institute of Sciences (Indonesia), the Forest Department Sarawak (Malaysia), Economic Planning Unit of the Prime Minister's Department (Malaysia), the Vietnam National Museum of Nature (Vietnam) and the Inland Fisheries Research and Development Institute (Cambodia). This publication has ISEM number SUD-xxx.

Data availability statement

DNA barcodes are available in the Barcode of Life Data System (dx.doi.org/110.5883/DS-BIFRA) and GenBank (see Table S3.1 for accession numbers). Mitogenomes are available in GenBank (see Table S3.1 for accession numbers).

Funding Information

This study was supported by IRD through annual allocations (2012-2020) and incentive funds (2012-2013), the Ministry of Foreign Affairs and International Development through a BIO-Asia grant (BIOSHOT 2016-2017) managed by Campus France, the “Institut Français d’Indonésie” through a “Science et impacts” grant (2016-5758) and the European Commission through a grant from the Southeast Asia – Europe Joint Funding scheme for Research and Innovation program (FRESHBIO, 307943/00). AS benefited from a PhD scholarship from the Indonesian Endowment Fund for Education (LPDP). The Indonesian Ministry of Research and Technology approved this study.

Chapter 4
Synthesis on Diversification of Sundaland Aquatic Biotas: Build-Up of Freshwater Fishes' Diversity and Distribution in a Biodiversity Hotspot

Freshwater Ichthyodiversity of Sundaland

The results of this PhD revealed significant level of cryptic freshwater ichthyodiversity within Sundaland biodiversity hotspot. Large ratios of molecular operational taxonomic unit (MOTU) to nominal species are detected, with ratios as much as 1.83 for *Clarias*, 2 for *Dermogenys* and *Glyptothorax*, 2.23 for Rasborinae, 3.17 for *Hemirhamphodon* and 3.6 for *Channa*. Considering that Sundaland hosts around 75% of known Indonesian freshwater ichthyodiversity, based on valid nominal species (Hubert et al. 2015b), this shows the importance of taking cryptic diversity into account. This result indicates cryptic diversity may massively escalate the expected total level of diversity. As a corollary, discounting this diversity will certainly lead to a major taxonomic bias that should be avoided for any further study and/or management applications (de Bruyn et al. 2004; Nguyen et al. 2008; Pouyaud et al. 2009; de Bruyn et al. 2013; Hubert et al. 2015a; Dahruddin et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018; Hubert et al. 2019). Method wise, our multiple approaches on the coalescent theory-based delimitation and alternative methods (Hutama et al. 2017; Hubert et al. 2019) have largely succeded in detetecting cryptic lineages (i.e. MOTU) and providing a robust timeframe of speciation for Sundaland freshwater fishes.

Following the apparent proportion of cryptic diversity recovered during the study, taxonomic works is critical on the study of Sundaland freshwater ichthyodiversity (Hubert et al. 2015b; Dahruddin et al. 2017; Hutama et al. 2017; Sholihah et al. 2020). Molecular taxonomic approaches based on coalescent theory used in this study has validated past taxonomic works as exemplified by: 1) lineage distinctiveness of *Brevibora exilis* Liao & Tan, 2014, *Clarias intermedius* Teugels, Sudarto & Pouyaud, 2001, *Clarias kapuasensis* Sudarto, Teugels & Pouyaud, 2003, *Clarias pseudoleiacanthus* Sudarto, Teugels & Pouyaud, 2003, *Clarias pseudonieuhofii* Sudarto, Teugels & Pouyaud, 2004, *Glyptothorax amnestus* Ng & Kottelat, 2016, *Glyptothorax exodon* Ng & Rachmatika, 2006, *Glyptothorax stibaros* Ng & Kottelat,

2016, *Glyptothorax pictus* Ng & Kottelat, 2016; *Rasbora maninjau* Lumbantobing, 2014, *Trigonostigma truncata* Tan, 2020, as well as 2) monophyly of genus *Boraras* Kottelat and Vidthayanon, 1993. Further and deeper taxonomic study are still highly needed, especially: 1) when cryptic lineages of a taxon do not constitute a monophyletic unit (e.g. *Clarias nieuhofii* (paraphyletic)*, Channa gachua* (paraphyletic)*, Glyptothorax fucsus* (paraphyletic), genus *Rasbora* (polyphyletic)); 2) when newly described species still contain cryptic diversity (e.g. *Rasbora patrickyapi*, Tan, 2009 – 3 MOTUs, and *Brevibora cheeya* Liao & Tan, 2011 – 2 MOTUs); or when newly described higher taxa are not monophyletic (e.g. genera *Trigonopoma* Liao, Kullander & Fang 2010 (paraphyletic) and *Brevibora* Liao, Kullander & Fang, 2010 (polyphyletic)). Particularly in the case of Rasborinae, phylogenetic reconstructions reveal that *Rasbora* species are scattered across all clades in Rasborinae phylogeny, a result in line with previous molecular and morphology-based phylogenetic studies, which highlighted that *Rasbora* is encompassing lineages, morphologically very similar, but of distinct evolutionary origins (Liao et al. 2010, 2011; Lumbantobing 2010; Tang et al. 2010; Tan and Armbruster 2018; Sholihah et al. 2020). This warrants further taxonomic works within Rasborinae and highlights the need of an in-depth revision of the genera *Rasbora* and its related newly descibed genera.

Diversification of Sundaland Aquatic Biotas

Biogeography of Sundaland Ichthyodiversity

Ancestral area reconstructions of *Clarias*, *Glyptothorax*, Zenarchopteridae, *Channa* and Rasborinae revealed that their Sundaland lineages originated from Mainland Asia. Even though evolved on different evolutionary timeframes, their colonisation events on Sundaland took place since Miocene for the first four taxa and as old as Oligocene for Rasborinae, thus generally reconfirm the previously suggested pre-Pleistocene settlement hypothesis of Sundaland (Dodson et al. 1995; Gorog et al. 2004). This hypothesis posits that, Borneo, the biggest emerging landmass of Sundaland, was still largely connected to the continent (Fig. 0.5), a situation that eased the onset of Sundaland colonization. Our reconstruction also match the geological reconstructions of Southeast Asia. The MRCA of Sundaland Rasborinae was infered to overlap with the onset of the isolation of Borneo from the continent; as well as by the matching between colonisation of Java (ca. 4 Ma the oldest) and the supposed age

of its modern geological setting (ca. 5 Myrs) (Hall 2009; Hall et al. 2011; Lohman et al. 2011).

Continental freshwater lineages are infered to enter Sundaland mainly through North Sunda palaeoriver in contemporary Borneo and later dispersed to other parts of Sundaland. As such, the Bornean part of North Sunda palaeoriver is the most likely centre of origin for the regional diversity of freshwater fishes. Beside, this area hosts a large proportion of Sundaland freshwater fishes diversity with Kapuas-Sentarum basin as its main stronghold. The watersheds in Borneo also include the northern part of East Sunda palaeoriver, northern Borneo and the Mahakam watershed, which share the same headwater in the centre of the island. Along the same line, East Sunda palaeoriver also shares boundaries in the lowlands, Southwestern Borneo, which might explains same lineages occur on both of palaeorivers. From this setting, it is expected that freshwater fishes from Bornean North Sunda palaeoriver dispersed to Sumatra via the same palaeoriver, however, headwater capture events among neighbouring, contiguous watersheds likely contributed to the spatial expansion of freshwater lineages in Sundaland.

Java was colonised the latest and has been subject to heavy level of freshwater habitat fragmentation due to its naturally fragmented geomorphology. *In situ* diversification within Javan freshwater habitats has been revealed in all analyses and all taxa studied here. Branches of East Sunda palaeoriver, the only palaeoriver flowing through this island, were cut into narrow and confined watershed that drove *in situ* diversification of freshwater fish lineages within them, creating local radiation which has also been reported before (Pouyaud et al. 2009; Kusuma et al. 2016; Hutama et al. 2017; Hubert et al. 2019). While this might have created higher level of species richness, it might also serve as a particular challenge on their survival within this contemporary time (Hutama et al. 2017; Hubert et al. 2019).

Biogeographic reconstruction in Sumatra involved much more complicated scenario than Java. Sumatra island is older than Java and also has volcanic activities unlike Borneo (Hall 2009; Hall et al. 2011; Lohman et al. 2011). Sumatra also has three palaeorivers flowing through it: Malacca strait (northern part), North Sunda (middle part), East Sunda (southern tip) (Voris 2000; Lohman et al. 2011). Generally though, it can be said that most of Sumatran lineages come through North Sunda and Malacca palaeorivers. The oldest recorded lineages in Rasborinae (late Oligocene to early Miocene) are associated to these palaeoriver systems and contrast with most lineages,

which are younger. This pattern suggests multiple colonization events either through Bornean North Sunda or directly from the continent via Malacca strait palaeoriver. While *in situ* diversification is, again, revealed from many Sumatran lineages, this island also shows several geographic area with unclear biogeographic affinity probably due to the existence of neighbouring palaeorivers flowing on the same lowlands, giving more chance for genetic flow (Hubert et al. 2015a), i.e. in Lampung between North Sunda and East Sunda palaeorivers.

Spatiotemporal Aspect of Diversification

Ancestral area estimations suggest that diversification of Sundaland freshwater fish lineages involved long distance dispersal as has been shown by the inclusion of J parameter (for jumping dispersal) on most likely models in all examined taxa. While settlement of their ancestral lineages generally took place before Sundaland was completely separated from the continent, suggesting that vicariance contributed to the initial steps of Sundaland ichthyodiversity build-up, the recent geological and palaeoenvironemental history likely favored long distance dispersal (de Bruyn et al. 2013; Condamine et al. 2015b; van Dam and Matzke 2016; Beck et al. 2017). The Palaeoriver Hypothesis, for instance, suggests that jump dispersal in freshwater fishes could be enabled during glacial period due to the emergence of temporary landbridges, which served as inter-island dispersal channels (Pouyaud et al. 2009; Adamson et al. 2012; de Bruyn et al. 2013).

Despite the apparent role of dispersal during the diversification of Sundaland freshwater fishes, the results of this study question the impact of the Pleistocene eustatic fluctuations on dispersal. Although lowered sea level during glacial periods reconnected watersheds within palaeorivers, it did not necessarily open up inter-island dispersal channels for freshwater fishes. As has been suggested before, cooler and arid climate enabled the interior of glacial Sundaland to develop corridor of savanna and seasonal forest ecosystems (Heaney 1991; Bird et al. 2005). These ecosystems are markedly different from the majority of ecosystems in interglacial periods. From this, we can expect that there would be barrier to dispersal for forest freshwater fishes, which are generally not observed in the dryer and more seasonal ecosystems. This result questions the effectiveness of available interisland freshwater channels for dispersal during glacial periods (Heaney 1991; Gathorne-Hardy et al. 2002; Gorog et

al. 2004; Bird et al. 2005; Pouyaud et al. 2009; Wurster et al. 2019). For examples, this has been shown in the diversification of peat swamp-dependent *Clarias* in Sundaland (Figs 1.4a and 1.5d) and many of the small size forested stream and/or peat swamp-dependent Rasborinae (Table 3.2, particularly from Clade I, II and III).

Pleistocene eustatic fluctuations created significant differences in the extent and connectivity of Sundaland palaeorivers. However, permeability of the physical boundaries of palaeoriver's watersheds are still arguable. Due to complex geomorphology of the region, margin between neighbouring palaeorivers might be porous such as when it covers generally flat, lowland freshwater habitats. As examples, boundaries between North Sunda and East Sunda palaeorivers in southern Sumatra and southwestern Borneo as well as between Malacca strait and North Sunda palaeorivers in Sumatra cannot be clearly delineated biologically as flat lowland might generate more chance for gene flow (Hubert et al. 2015a). On the contrary, allopartic speciation might happen internally when varied coexisting ecosystems or geomorphological complexity enable *in situ* diversification within the same palaeoriver. Largely exemplified in Java, complex geological setting has fragmented East Sunda palaeoriver in the island and drove local radiations (Kusuma et al. 2016; Hutama et al. 2017; Hubert et al. 2019). Such cases were detected in *Channa gachua, Dermogenys pusilla* as well as several species of *Glyptothorax* and Rasborinae Clade IV. Besides, each Sundaland palaeoriver covers different types of ecosystem, both during glacial and interglacial periods. For example, contemporary East Sunda palaeoriver encompasses tropical rainforest, peat swamp, and floodplain freshwater habitats in Borneo while peat swamp habitats are mostly absent in Java. During glacial period, savanna and seasonal forest associated freshwater basin might also add up to this composition (Heaney 1991; Bird et al. 2005). Considering this, it will not be a surprise to see that many of Sundaland freshwater taxa will be specific to certain type of freshwater habitats despite occupying the same palaeoriver. Such cases of specialization are detected in *Clarias* with blackwater versus non-blackwater lineages, or blackwater versus riverine Rasborinae.

Temporal wise, Pleistocene eustatic fluctuation and Pleistocene palaeoriver dynamics did not elevate diversification rate of Sundaland freshwater fishes as previously expected (Kottelat et al. 1993; Voris 2000; Mittelbach et al. 2007; Woodruff 2010; Condamine et al. 2015a; Li and Li 2018). The only case of sea-level related diversification was observed in the genus *Channa*, which is highly opportunistic and

easy to disperse (in particular *Channa striata*) (Fig 1.3). Considering both dispersal and allopatric speciation driven by palaeoriver dynamics have been suggested to be predominant, this finding is novel and surprising. Along the same line, none of the examined taxa under this study, which has substantial proportion of Pleistocene lineages, has declining diversification rates as suggested by diversity-dependent diversification (DDD) model (Seehausen 2007; Rabosky and Lovette 2008; Seehausen et al. 2008). Constant birth (BCST) models of diversification are infered to be most likely for almost all examined groups. In this case, rather than resulting from elevated speciation and bounded by the saturation of environmental carrying capacity (Schluter 2000; Hubbell 2001; Alonso et al. 2006; Phillimore and Price 2008; Kisel et al. 2011), large proportion of Pleistocene lineages is resulted from ongoing constant speciation through time. The speciation rates (λ) f inferred here for *Clarias*, *Glyptothorax* and Rasborinae Clade IV are respectively 0.4556, 0.5444, 0.1548 lineages/Myr. Compared with the 0.141 lineages/Myr mean rate of speciation for fishes (Rabosky et al. 2013), these speciation rates are high. This rate is even higher for Zenarchopteridae with an initial speciation rate λ_0 = 0.432 and a progressive increase through time. The heterogeneity of molecular markers used likely account for the differences observed between speciation rates estimated in chapter 1 and chapter 3. Mitogenomes proved to recover highly supported phylogenetic hypotheses of relationships and accurate age estimates, particularly for ancient events (Saitoh Miya, M., Inoue, J. G., Ishiguro, N. B., Nishida, M. 2003; Lavoué Miya, M., Inoue, J. G., Saitoh, K., Ishiguro, N. B., Nishida, M. 2005; Saitoh et al. 2006; Lavoué Miya, M, Saitoh, K, Ishiguro, NB, Nishida, M 2007; Kawahara et al. 2008). The speciation rates inferred for Rasborinae clades seems to confirm the usefulness of mitogenomes for macroevolutionary studies as they corroborate speciation rates inferred from large scale fish phylogenetic trees (Rabosky et al. 2013) Alternatively, speciation rate estimates in Chapter 1 are based on multi-locus approaches, consisting of multiple, short fragments, and more prone to bias in age estimates than mitogenomes due to homoplasy. This discrepancy warrants further studies on Clariidae, Channidae, *Glyptothorax* and Beloniformes speciation rates through more comprehensive genomic sampling.

Key Aspects for Generating Sundaland Freshwater Ichthyodiversity

Sundaland freshwater fish communities originated from continental Asia, came (mainly) through Bornean North-Sunda palaeoriver since Oligocene, then underwent a pre-Pleistocene diversification following tectonic movement of Borneo, as well as the later internal insular vicariance by the opening up of regional shallow seas around Sundaland main landmasses. The initial lineages of freshwater fishes in Sundaland proliferated through time, driven by biotic and abiotic factors, both on global/regional and local scales. Contrary to Pleistocene palaeoriver hypothesis though, our findings suggest that rather than soliciting global Pleistocene eustatic fluctuation and regional palaeoriver dynamics as drivers for Sundaland freshwater fish diversification, we should actually consider to reconcile them with local scale heterogeneity (Haffer 1997; Nores 1999; Brown et al. 2013; Hubert et al. 2015a). And, by considering that Sundaland freshwater basins are large and cover highly heterogenous types of habitat, we might need to emphasize that ecological factors probably had a more important role in maintaining this staggering diversity than previously expected (Mittelbach et al. 2007). Meanwhile, the impact of global Pleistocene eustatic fluctuation and regional palaeoriver dynamics might still exist but possibly much more limited to taxa with particular life history traits. In this context though, taxa with high adaptability and high dispersal ability are expected to thrive successfully and adapt to ecological heterogeneity (Mittelbach et al. 2007). This is also in line with the previosly suggested scenario, that although generally Pleistocene climatic fluctuation (abiotic factor) affected greatly the evolution of temperate organisms, natural selection in tropical system relies more on biotic interactions, favouring different adaptation toward specific niche/habitat (Mittelbach et al. 2007). Table 5.1 below shows how the global Pleistocene eustatic is actually just a part of abiotic aspects affecting the diversification of Sundaland freshwater fishes.

Table 4.1 Key Aspects in Sundaland Freshwater Fish Diversification

	Abiotic		Biotic	
	Global Pleistocene Eustatic	**Local Geomorphology**	**Local Ecosystem/ Habitat Variability**	**Life History Traits**
Between Palaeorivers	• Contraction or expansion of palaeoriver boundaries	• Porous palaeoriver boundaries • Dispersal inter palaeorivers (within and/or between islands)	• Presence of specific habitat in land-bridges • Composition of ecological communities in land-bridges and competition	• Taxa with high dispersal ability might exploit porous palaeoriver boundaries
Within Palaeoriver	• Allopatric speciation within palaeoriver on different islands during interglacial period • Dispersal within palaeoriver on different islands during glacial period	• Internal palaeoriver fragmentation into smaller and/or confined freshwater habitats • Allopatric speciation within palaeoriver	• Internal palaeoriver fragmentation into smaller and/or confined freshwater habitats • Ecological speciation within palaeoriver • Barrier for dispersal during glacial period due to the lack of suitable habitat	• Taxa with high dispersal ability might exploit glacial palaeoriver channel despite different biotic and abiotic conditions • Given specific ecosystem and/or geological setting, taxa with specific niches/habitat requirements might undergo local radiation within palaeoriver

Pespectives and Implications

Considering only the physical extent of freshwater basin, contemporary freshwater habitats in Sundaland are indeed in refugial state (Voris 2000; Woodruff 2010; Lohman et al. 2011). Even so, if we take into account the most likely types of freshwater habitats and their associated terrestrial ecosystems, which existed during past glacial periods (Heaney 1991; Bird et al. 2005; Wurster et al. 2019), current freshwater habitats have been more or less stable through time. For this, it is important to emphasize that based on present studies, Sundaland freshwater fish diversity is rich, probably not because of the isolation and reconnection mechanism during Pleistocene, but rather by long term stability, both geologically and ecologically, facilitating high and persisting species proliferation (Haffer 1997; Gathorne-Hardy et

al. 2002; Gorog et al. 2004). Given this, conservation of Sundaland freshwater habitats is definitely the key for survival of Sundaland freshwater biodiversity as a whole.

Several practical points can be generated from this study for facilitating conservation efforts in the region. **First**, although it is not generally the main driver for diversification, we still see concordance in species composition within the same palaeoriver, either within the same or different islands. For this, rather than only leaning conservation management of freshwater habitats towards its terrestrial counterpart, it is also important to recognize more biogeography-oriented management strategies by clustering them based on palaeoriver & island-based consideration. **Second**, it is important to recognize MOTUs, instead of simply nominal species, while identifying biodiversity richness of freshwater habitats. As have been stated several times before, Sundaland freshwater fishes have significant proportion of cryptic diversity with multiple, highly divergent lineages (de Bruyn et al. 2004; Nguyen et al. 2008; Pouyaud et al. 2009; de Bruyn et al. 2013; Hubert et al. 2015a, 2015b; Dahruddin et al. 2017; Hutama et al. 2017; Nurul Farhana et al. 2018; Hubert et al. 2019; Sholihah et al. 2020). Therefore, recognizing this cryptic diversity is important both to avoid underestimating the level of biodiversity in Sundaland as well as to conduct better and productive conservation management as a whole, i.e. during translocation and/or reintroduction of lineages or when contemporary demographic trend is needed for infering the Maximum Sustainable Yield (MSY) of commercial freshwater catching fishery (Russell 1931; Graham 1935; Holden and Ellner 2016; Hutama et al. 2017). Recognition of MOTUs can also facilitate better decision making in a commonly used “species approach” conservation efforts. For example, as mandated by Indonesian law on the conservation of plants and animal (Article 5 Government Regulation Number 7/1999), any kind of animal with small population size or limited distribution (endemic) should be given status “protected”, which both are exactly the characteristics of most MOTUs revealed in this study. **Third**, given that Sundaland freshwater fishes’ MOTUs are mostly restricted to particular geograhical/ecosystem range and many of them are confined within a very narrow habitat setting (Pouyaud et al. 2009; Dahruddin et al. 2017; Hutama et al. 2017; Hubert et al. 2019), it is important for freshwater conservation in Sundaland to take into account not only the extent of protected freshwater habitat, but also its ecological and geological complexities. **Fourth**, while conservation efforts for freshwater habitats in Sumatra and Borneo have been more profound, for example by designation of Ramsar sites in Berbak and Sentarum, Java

which has high level of *in situ* diversification and cryptic diversity despite generally lower total species richness also need to be the focus. Different with the other two islands, Java surely has much more degraded freshwater ecosystems, which are confined both by its geological nature and contemporary anthropogenic footprints, calling for a distinct conservation strategy. Rather than heavily depending on designating particular freshwater habitats as protected areas, conservation efforts in Java can be facilitated by general environmental rehabilitation, such as: 1) rehabilitation of water catchment areas; 2) application of appropriate and ecologically friendly agriculture; 3) appropriate wastewater and general waste management systems; 4) regulated fishing; etc. Furthermore, when designating protected freshwater areas in Java is possible, it is important to identify molecular lineages and distribution of the existing freshwater fishes in the area, which can be facilitated either by refering to present studies and previous studies on Javanese ichthyodiversity as well as by utilizing international repositories such as BOLD system (Ratnasingham and Hebert 2007, 2013; Hubert et al. 2015b, 2019; Dahruddin et al. 2017; Hutama et al. 2017; Sholihah et al. 2020).

References

Adamson E.A.S., Hurwood D.A., Mather P.B. 2010. A reappraisal of the evolution of Asian snakehead fishes (Pisces, Channidae) using molecular data from multiple genes and fossil calibration. Mol. Phylogenet. Evol. 56:707–717.

Adamson E.A.S., Hurwood D.A., Mather P.B. 2012. Insights into historical drainage evolution based on the phylogeography of the chevron snakehead fish (*Channa striata*) in the Mekong Basin. Freshw. Biol. 57:2211–2229.

Alonso D., Etienne R., McKane A. 2006. The merits of neutral theory. Trends Ecol. Evol. 21:451–457.

Alter S.E., Munshi-South J., Stiassny M.L.J. 2017. Genomewide SNP data reveal cryptic phylogeographic structure and microallopatric divergence in a rapids-adapted clade of cichlids from the Congo River. Mol. Ecol. 26:1401–1419.

April J., Hanner R.H., Mayden R.L., Bernatchez L. 2013. Metabolic rate and climatic fluctuations shape continental wide pattern of genetic divergence and biodiversity in fishes. PLoS One. 8:e70296.

April J., Mayden L.R., Hanner R.H., Bernatchez L. 2011. Genetic calibration of species diversity among North America's freshwater fishes. Proc. Natl. Acad. Sci. USA. 108.

Arita H.T., Vázquez-Domínguez E. 2008. The tropics: cradle, museum or casino? A dynamic null model for latitudinal gradients of species diversity. Ecol. Lett. 11:653–663.

Avise J.C. 1994. Molecular markers, natural history and evolution. Boston, MA: Springer US.

Avise J.C. 2000. Phylogeography: The history and formation of species. Cambridge: Harvard University Press.

Avise J.C. 2009. Phylogeography: retrospect and prospect. J. Biogeogr. 36:3–15.

Barluenga M., Meyer A. 2004. The Midas cichlid species complex: Incipient sympatric speciation in Nicaraguan cichlid fishes? Mol. Ecol. 13:2061–2076.

Barraclough T.G., Nee S. 2001. Phylogenetics and speciation. Trends Ecol. Evol. 16:391–399.

Bates M., Hackett, S., J., Cracraft, J. J. 1998. Area-relationships in the neotropical lowlands: an hypothesis based on raw distributions of passerine birds. J. Biogeogr. 25:783–793.

Beck S. V., Carvalho G.R., Barlow A., Ruber L., Tan H.H., Nugroho E., Wowor D., Mohd Nor S.A., Herder F., Muchlisin Z.A., de Bruyn M. 2017. Plio-Pleistocene phylogeography of the Southeast Asian Blue Panchax killifish, *Aplocheilus panchax*. PLoS One. 12:e0179557.

Bermingham E., McCafferty S.S., Martin A.P. 1997. Fish biogeography and molecular clocks: perspectives from the Panamanian Isthmus. In: Kocher T.D., Stepien C.A., editors. Molecular systematics of fishes. Elsevier. p. 113–128.

Bernatchez L., Wilson C. 1998. Comparative phylogography of Neartic and Paleartic fishes. Mol. Ecol. 7:431–452.

Berra T.M. 2001. Freshwater fish distribution. San Diego: Academic Press.

Betancur-R R., Wiley E.O., Arratia G., Acero A., Bailly N., Miya M., Lecointre G., Ortí G. 2017. Phylogenetic classification of bony fishes. BMC Evol. Biol. 17:162.

Bird M.I., Taylor D., Hunt C. 2005. Palaeoenvironments of insular Southeast Asia

during the Last Glacial Period: A savanna corridor in Sundaland? Quat. Sci. Rev. 24:2228–2242.
Blagoev G.A., deWaard J.R., Ratnasingham S., deWaard S.L., Lu L., Robertson J., Telfer A.C., Hebert P.D.N. 2016. Untangling taxonomy: A DNA barcode reference library for Canadian spiders. Mol. Ecol. Resour. 16:325–341.
Blair C., Bryson R.W. 2017. Cryptic diversity and discordance in single-locus species delimitation methods within horned lizards (Phrynosomatidae: *Phrynosoma*). Mol. Ecol. Resour. 17:1168–1182.
Blaxter M., Mann J., Chapman T., Thomas F., Whitton C., Floyd R., Abebe E. 2005. Defining operational taxonomic units using DNA barcode data. Philos. Trans. R. Soc. B Biol. Sci. 360:1935–1943.
Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10:e1003537.
Brittan M.R. 1954. A revision of the Indo-Malayan frash-water fish genus *Rasbora*. Monogr. Inst. Sci. Tech. Manila. 3:3 pls.
Brittan M.R. 1972. Rasbora: A revision of the Indo-Malayan freshwater fish genus *Rasbora*. Hongkong: TFH Publications.
Brown R.M., Siler C.D., Oliveros C.H., Esselstyn J.A., Diesmos A.C., Hosner P.A., Linkem C.W., Barley A.J., Oaks J.R., Sanguila M.B., Welton L.J., Blackburn D.C., Moyle R.G., Townsend Peterson A., Alcala A.C. 2013. Evolutionary processes of diversification in a model island archipelago. Annu. Rev. Ecol. Evol. Syst. 44:411–435.
Brown S.D.J., Collins R.A., Boyer S., Lefort M.C., Malumbres-Olarte J., Vink C.J., Cruickshank R.H. 2012. Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Mol. Ecol. Resour. 12:562–565.
de Bruyn M., Rüber L., Nylinder S., Stelbrink B., Lovejoy N.R., Lavoué S., Tan H.H., Nugroho E., Wowor D., Ng P.K.L., Siti Azizah M.N., von Rintelen T., Hall R., Carvalho G.R. 2013. Paleo-drainage basin connectivity predicts evolutionary relationships across three Southeast Asian biodiversity hotspots. Syst. Biol. 62:398–410.
de Bruyn M., Stelbrink B., Morley R.J., Hall R., Carvalho G.R., Cannon C.H., van den Bergh G., Meijaard E., Metcalfe I., Boitani L., Maiorano L., Shoup R., von Rintelen T. 2014. Borneo and Indochina are major evolutionary hotspots for Southeast Asian biodiversity. Syst Biol. 63:879–901.
de Bruyn M., Wilson J.A., Mather P.B. 2004. Huxley's line demarcates extensive genetic divergence between eastern and western forms of the giant freshwater prawn, *Macrobrachium rosenbergii*. Mol. Phylogenet. Evol. 30:251–257.
Burbrink F.T., Chan Y.L., Myers E.A., Ruane S., Smith B.T., Hickerson M.J. 2016. Asynchronous demographic responses to Pleistocene climate change in Eastern Nearctic vertebrates. Ecol. Lett. 19:1457–1467.
Cannon C.H., Morley R.J., Bush A.B. 2009. The current refugial rainforests of Sundaland are unrepresentative of their biogeographic past and highly vulnerable to disturbance. Proc Natl Acad Sci U S A. 106:11188–11193.
Clark J.R., Ree R.H., Alfaro M.E., King M.G., Wagner W.L., Roalson E.H. 2008. A comparative study in ancestral range reconstruction methods: Retracing the uncertain histories of insular lineages. Syst. Biol. 57:693–707.
Clements R., Sodhi N.S., Schilthuizen M., Ng P.K.L. 2006. Limestone karsts of Southeast Asia: Imperiled arks of biodiversity. Bioscience. 56:733–742.

Collins R.A., Armstrong K.F., Meier R., Yi Y., Brown S.D.J., Cruickshank R.H., Keeling S., Johnston C. 2012. Barcoding and border biosecurity: Identifying cyprinid fishes in the aquarium trade. PLoS One. 7.
Condamine F.L., Nagalingum N.S., Marshall C.R., Morlon H. 2015a. Origin and diversification of living cycads: a cautionary tale on the impact of the branching process prior in Bayesian molecular dating. BMC Evol. Biol. 15:65.
Condamine F.L., Rolland J., Morlon H. 2013a. Macroevolutionary perspectives to environmental change. Ecol. Lett. 16:72–85.
Condamine F.L., Rolland J., Morlon H. 2019. Assessing the causes of diversification slowdowns: Temperature-dependent and diversity-dependent models receive equivalent support. Ecol. Lett. 22:1900–1912.
Condamine F.L., Toussaint E.F.A., Clamens A.-L., Genson G., Sperling F.A.H., Kergoat G.J. 2015b. Deciphering the evolution of birdwing butterflies 150 years after Alfred Russel Wallace. Sci. Rep. 5:11860.
Condamine F.L., Toussaint E.F.A., Cotton A.M., Genson G.S., Sperling F.A.H., Kergoat G.J. 2013b. Fine-scale biogeographical and temporal diversification processes of peacock swallowtails (*Papilio* subgenus *Achillides*) in the Indo-Australian Archipelago. Cladistics. 29:88–111.
Conte-Grand C., Britz R., Dahanukar N., Raghavan R., Pethiyagoda R., Tan H.H., Hadiaty R.K., Yaakob N.S., Rüber L. 2017. Barcoding snakeheads (Teleostei, Channidae) revisited: Discovering greater species diversity and resolving perpetuated taxonomic confusions. PLoS One. 12:e0184017.
Conway K.W., Chen W.J., Mayden R.L. 2008. The “celestial pearl danio” is a miniature *Danio* (s.s) (Ostariophysi: Cyprinidae): Evidence from morphology and molecules. Zootaxa. 1686:1–28.
Conway K.W., Hirt M.V., Yang L., Mayden R.L., Simons A.M. 2010. Cypriniformes: Systematics & paleontology: Festschrift in honor of G. Arratia. Origin and Phylogenetic Interrelationships of Teleosts. p. 295–316.
Cornell H. V. 1993. Unsaturated patterns in species assemblages: the role of regional processes in setting local species richness. In: Ricklefs Schluter, D R.E., editor. Species diversity in ecological communities: historical and geographical perspectives. Chicago: University of Chicago Press. p. 243–252.
Cowie R.H., Holland B.S. 2006. Dispersal is fundamental to biogeography and the evolution of biodiversity on oceanic islands. J. Biogeogr. 33:193–198.
Currie D.J., Mittelbach G.G., Cornell H. V., Field R., Guegan J.-F., Hawkins B.A., Kaufman D.M., Kerr J.T., Oberdorff T., O’Brien E., Turner J.R.G. 2004. Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. Ecol. Lett. 7:1121–1134.
Dahruddin H., Hutama A., Busson F., Sauri S., Hanner R., Keith P., Hadiaty R., Hubert N. 2017. Revisiting the ichthyodiversity of Java and Bali through DNA barcodes: taxonomic coverage, identification accuracy, cryptic diversity and identification of exotic species. Mol. Ecol. Resour. 17:288–299.
van Dam M.H., Matzke N.J. 2016. Evaluating the influence of connectivity and distance on biogeographical patterns in the south-western deserts of North America. J. Biogeogr. 43:1514–1532.
Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: More models, new heuristics and parallel computing. Nat. Methods. 9:772.
Dodson J.J., Colombani F., Ng P.K.L. 1995. Phylogeographic structure in mitochondrial DNA of a South-east Asian freshwater fish, *Hemibagrus nemurus* (Siluroidei; Bagridae) and Pleistocene sea-level changes on the Sunda shelf. Mol.

Ecol. 4:331–346.
Dodsworth S. 2015. Genome skimming for next-generation biodiversity analysis. Trends Plant Sci. 20:525–527.
Dong J., Kergoat G.J., Vicente N., Rahmadi C., Xu S., Robillard T. 2018. Biogeographic patterns and diversification dynamics of the genus *Cardiodactylus* Saussure (Orthoptera, Grylloidea, Eneopterinae) in Southeast Asia. Mol Phylogenet Evol. 129:1–14.
Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.
Duchêne S., Lanfear R., Ho S.Y.W. 2014. The impact of calibration and clock-model choice on molecular estimates of divergence times. Mol. Phylogenet. Evol. 78:277–289.
Durand J., Persat H., Bouvet Y. 1999. Phylogeography and postglacial dispersion of the chub (*Leuciscus cephalus*) in Europe. Mol. Ecol. 8:989–997.
Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Eschmeyer W.N., Fricke R., van der Laan R. 2018. Catalog of fishes electronic version.
Esselstyn J.A., Brown R.M. 2009. The role of repeated sea-level fluctuations in the generation of shrew (Soricidae: *Crocidura*) diversity in the Philippine Archipelago. Mol. Phylogenet. Evol. 53:171–181.
Esselstyn J.A., Timm R.M., Brown R.M. 2009. Do geological or climatic processes drive speciation in dynamic archipelagos? The tempo and mode of diversification in Southeast Asian shrews. Evolution (N. Y). 63:2595–2610.
Etienne R.S., Haegeman B., Stadler T., Aze T., Pearson P.N., Purvis A., Phillimore A.B. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proc. R. Soc. B Biol. Sci. 279:1300–1309.
Fang F., Norén M., Liao T.Y., Källersjö M., Kullander S.O. 2009. Molecular phylogenetic interrelationships of the south Asian cyprinid genera *Danio*, *Devario* and *Microrasbora* (Teleostei, Cyprinidae, Danioninae). Zool. Scr. 38:237–256.
Froese R., Pauly D. 2014. Fishbase. Available from http://www.fishbase.org.
Froese R., Pauly D. 2020. Fishbase. Available from http://www.fishbase.org.
Fujisawa T., Barraclough T.G. 2013. Delimiting species using single-locus data and the generalized mixed yule coalescent approach: A revised method and evaluation on simulated data sets. Syst. Biol. 62:707–724.
Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397–423.
Gaston K.J., Spicer J.I. 2004. Biodiversity: An introduction. Oxford: Blackwell Publishing Ltd.
Gathorne-Hardy F.J., Syaukani, Davies R.G., Eggleton P., Jones D.T. 2002. Quaternary rainforest refugia in south-east Asia: using termites (Isoptera) as indicators. Biol. J. Linn. Soc. 75:453–466.
Gavrilets S., Vose A. 2006. Dynamic patterns of adaptative radiation. Proc. Natl. Acad. Sci. USA. 102:18040–18045.
Giam X., Koh L.P., Tan H.H., Miettinen J., Tan H.T.W., Ng P.K.L. 2012. Global extinctions of freshwater fishes follow peatland conversion in Sundaland. Front. Ecol. Environ. 10:465–470.
Gorog A.J., Sinaga M.H., Engstrom M.D. 2004. Vicariance or dispersal? Historical biogeography of three Sunda shelf marine rodents (*Maxomys surifer*, *Leopoldamys sabanus* and *Maxomys whiteheadi*). Biol. J. Linn. Soc. 81:91–109.

Graham M. 1935. Modern theory of exploiting a fishery, and application to north sea trawling. ICES J. Mar. Sci. 10:264–274.
Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.
Guyot J.L. 1993. Hydrogéochimie des fleuves de L'Amazonie Boliviene. .
Haffer J. 1997. Alternative models of vertebrate speciation in Amazonia: An overview. Biodivers. Conserv. 6:451–476.
Hall R. 2009. Southeast Asia's changing palaeogeography. Blumea J. Plant Taxon. Plant Geogr. 54:148–161.
Hall R. 2012. Late Jurassic-Cenozoic reconstructions of the Indonesian region and the Indian Ocean. Tectonophysics. 570–571:1–41.
Hall R. 2013. The palaeogeography of Sundaland and Wallacea since the Late Jurassic. J. Limnol. 72:1–17.
Hall R., Cottam M.A., Wilson M.E.J. 2011. The SE Asian gateway: history and tectonics of the Australia–Asia collision. Geol. Soc. London, Spec. Publ. 355:1–6.
Hardman M., Lundberg J.G. 2006. Molecular phylogeny and a chronology of diversification for "phractocephaline" catfishes (Siluriformes: Pimelodidae) based on mitochondrial DNA and nuclear recombination activating gene 2 sequences. Mol. Phylogenet. Evol. 40:410–418.
Heaney L.R. 1991. A synopsis of climatic and vegetational change in Southeast Asia. Clim. Change. 19:53–61.
Heaney L.R. 2007. Is a new paradigm emerging for oceanic island biogeography? J. Biogeogr. 34:753–757.
Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Hendriks K.P., Alciatore G., Schilthuizen M., Etienne R.S. 2019. Phylogeography of Bornean land snails suggests long-distance dispersal as a cause of endemism. J. Biogeogr. 46:932–944.
Herder F., Nolte A.W., Pfaender J., Schwarzer J., Hadiaty R.K., Schliewen U.K. 2006. Adaptive radiation and hybridization in Wallace's Dreamponds: evidence from sailfin silversides in the Malili Lakes of Sulawesi. Proc. R. Soc. B Biol. Sci. 273:2209–2217.
Hirt M.V., Arratia G., Chen W.J., Mayden R.L., Tang K.L., Wood R.M., Simons A.M. 2017. Effects of gene choice, base composition and rate heterogeneity on inference and estimates of divergence times in cypriniform fishes. Biol. J. Linn. Soc. 121:319–339.
Ho S.Y.W., Larson G. 2006. Molecular clocks: when timesare a-changin'. Trends Genet. 22:79–83.
Hoffmann M., Hilton-Taylor C., Angulo A., Böhm M., Brooks T.M., Butchart S.H.M., Carpenter K.E., Chanson J., Collen B., Cox N.A., Darwall W.R.T., Dulvy N.K., Harrison L.R., Katariya V., Pollock C.M., Quader S., Richman N.I., Rodrigues A.S.L., Tognelli M.F., Vié J.C., Aguiar J.M., Allen D.J., Allen G.R., Amori G., Ananjeva N.B., Andreone F., Andrew P., Ortiz A.L.A., Baillie J.E.M., Baldi R., Bell B.D., Biju S.D., Bird J.P., Black-Decima P., Blanc J.J., Bolaños F., Bolivar-G. W., Burfield I.J., Burton J.A., Capper D.R., Castro F., Catullo G., Cavanagh R.D., Channing A., Chao N.L., Chenery A.M., Chiozza F., Clausnitzer V., Collar N.J., Collett L.C., Collette B.B., Cortez Fernandez C.F., Craig M.T., Crosby M.J., Cumberlidge N., Cuttelod A., Derocher A.E., Diesmos A.C., Donaldson J.S., Duckworth J.W., Dutson G., Dutta S.K., Emslie R.H., Farjon A., Fowler S., Freyhof J., Garshelis D.L., Gerlach J., Gower D.J., Grant T.D., Hammerson G.A., Harris

R.B., Heaney L.R., Hedges S.B., Hero J.M., Hughes B., Hussain S.A., Icochea M. J., Inger R.F., Ishii N., Iskandar D.T., Jenkins R.K.B., Kaneko Y., Kottelat M., Kovacs K.M., Kuzmin S.L., La Marca E., Lamoreux J.F., Lau M.W.N., Lavilla E.O., Leus K., Lewison R.L., Lichtenstein G., Livingstone S.R., Lukoschek V., Mallon D.P., McGowan P.J.K., McIvor A., Moehlman P.D., Molur S., Alonso A.M., Musick J.A., Nowell K., Nussbaum R.A., Olech W., Orlov N.L., Papenfuss T.J., Parra-Olea G., Perrin W.F., Polidoro B.A., Pourkazemi M., Racey P.A., Ragle J.S., Ram M., Rathbun G., Reynolds R.P., Rhodin A.G.J.., Richards S.J., Rodríguez L.O., Ron S.R., Rondinini C., Rylands A.B., de Mitcheson Y.S., Sanciangco J.C., Sanders K.L., Santos-Barrera G., Schipper J., Self-Sullivan C., Shi Y., Shoemaker A., Short F.T., Sillero-Zubiri C., Silvano D.L., Smith K.G., Smith A.T., Snoeks J., Stattersfield A.J., Symes A.J., Taber A.B., Talukdar B.K., Temple H.J., Timmins R., Tobias J.A., Tsytsulina K., Tweddle D., Ubeda C., Valenti S. V., Van Dijk P.P., Veiga L.M., Veloso A., Wege D.C., Wilkinson M., Williamson E.A., Xie F., Young B.E., Akçakaya H.R., Bennun L., Blackburn T.M., Boitani L., Dublin H.T., da Fonseca G.A.B., Gascon C., Lacher T.E., Mace G.M., Mainka S.A., McNeely J.A., Mittermeier R.A., Reid G.M.G., Rodriguez J.P., Rosenberg A.A., Samways M.J., Smart J., Stein B.A., Stuart S.N. 2010. The impact of conservation on the status of the world's vertebrates. Science (80-.). 330:1503–1509.

Holden M.H., Ellner S.P. 2016. Human judgment vs. quantitative models for the management of ecological resources. Ecol. Appl. 26:1553–1565.

Hubbell S.P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton: Princeton University Press.

Hubert N., Calcagno V., Etienne R.S., Mouquet N. 2015a. Metacommunity speciation models and their implications for diversification theory. Ecol. Lett. 18:864–881.

Hubert N., Dettai A., Pruvost P., Cruaud C., Kulbicki M., Myers R., Borsa P. 2017. Geography and life history traits account for the accumulation of cryptic diversity among Indo-West Pacific coral reef fishes. Mar. Ecol. Prog. Ser. 583:179–193.

Hubert N., Duponchelle F., Nuñez J., Garcia-Davila C., Paugy D., Renno J.F. 2007. Phylogeography of the piranha genera *Serrasalmus* and *Pygocentrus*: Implications for the diversification of the Neotropical ichthyofauna. Mol. Ecol. 16:2115–2136.

Hubert N., Hanner R. 2015. DNA barcoding, species delineation and taxonomy: A historical perspective. DNA Barcodes. 3:44–58.

Hubert N., Hanner R., Holm E., Mandrak N.E., Taylor E., Burridge M., Watkinson D., Dumont P., Curry A., Bentzen P., Zhang J., April J., Bernatchez L. 2008. Identifying Canadian freshwater fishes through DNA barcodes. PLoS One. 3:e2490.

Hubert N., Kadarusman, Wibowo A., Busson F., Caruso D., Sulandari S., Nafiqoh N., Pouyaud L., Rüber L., Avarre J.-C., Herder F., Hanner R., Keith P., Hadiaty R.K. 2015b. DNA barcoding Indonesian freshwater fishes: Challenges and prospects. DNA Barcodes. 3:144–169.

Hubert N., Lumbantobing D., Sholihah A., Dahruddin H., Delrieu-Trottin E., Busson F., Sauri S., Hadiaty R., Keith P. 2019. Revisiting species boundaries and distribution ranges of *Nemacheilus spp.* (Cypriniformes: Nemacheilidae) and *Rasbora spp.* (Cypriniformes: Cyprinidae) in Java, Bali and Lombok through DNA barcodes: Implications for conservation in a biodiversity hotspot. Conserv. Genet. 20:517–529.

Hubert N., Meyer C.P., Bruggemann H.J., Guérin F., Komeno R.J.L., Espiau B., Causse R., Williams J.T., Planes S. 2012. Cryptic diversity in Indo-Pacific coral-

reef fishes revealed by DNA-barcoding provides new support to the centre-of-overlap hypothesis. PLoS One. 7:e28987.
Hubert N., Renno J.-F. 2006. Historical biogeography of South American freshwater fishes. J. Biogeogr. 33:1414–1436.
Husson L., Boucher F.C., Sarr A., Sepulchre P., Cahyarini S.Y. 2019. Evidence of Sundaland's subsidence requires revisiting its biogeography. J. Biogeogr.:jbi.13762.
Hutama A., Dahruddin H., Busson F., Sauri S., Keith P., Hadiaty R.K., Hanner R., Suryobroto B., Hubert N. 2017. Identifying spatially concordant evolutionary significant units across multiple species through DNA barcodes: Application to the conservation genetics of the freshwater fishes of Java and Bali. Glob. Ecol. Conserv. 12:170–187.
Ivanova N. V, Zemlak T.S., Hanner R.H., Hébert P.D.N. 2007. Universal primers cocktails for fish DNA barcoding. Mol. Ecol. Notes. 7:544–548.
Iwasaki W., Fukunaga T., Isagozawa R., Yamada K., Maeda Y., Satoh T.P., Sado T., Mabuchi K., Takeshima H., Miya M., Nishida M. 2013. Mitofish and mitoannotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol. Biol. Evol. 30:2531–2540.
Jiang W., Ng H.H., Yang J., Chen X. 2011. Monophyly and phylogenetic relationships of the catfish genus *Glyptothorax* (Teleostei: Sisoridae) inferred from nuclear and mitochondrial gene sequences. Mol. Phylogenet. Evol. 61:278–289.
Kadarusman, Hubert N., Hadiaty R.K., Sudarto, Paradis E., Pouyaud L. 2012. Cryptic diversity in Indo-Australian rainbowfishes revealed by DNA barcoding: Implications for conservation in a biodiversity hotspot candidate. PLoS One. 7:e40627.
Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587–589.
Kapli P., Lutteropp S., Zhang J., Kobert K., Pavlidis P., Stamatakis A., Flouri T. 2017. Multi-rate Poisson Tree Processes for single-locus species delimitation under Maximum Likelihood and Markov Chain Monte Carlo. Bioinformatics. 33:btx025.
Kawahara R., Miya M., Mabuchi K., Lavoué S., Inoue J.G., Satoh T.P., Kawaguchi A., Nishida M. 2008. Interrelationships of the 11 Gasterosteiform families (sticklebacks, pipefishes, and their relatives): A new perspective based on whole mitogenome sequences from 75 higher Teleosts. Mol. Phylogenet. Evol. 46:224–236.
Keith P., Lord C., Darhuddin H., Limmon G., Sukmono T., Hadiaty R., Hubert N. 2017. *Schismatogobius* (Gobiidae) from Indonesia, with description of four new species. Cybium. 41:195–211.
Kekkonen M., Hebert P.D.N. 2014. DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Mol. Ecol. Resour. 14:706–715.
Kekkonen M., Mutanen M., Kaila L., Nieminen M., Hebert P.D.N. 2015. Delineating species with DNA barcodes: A case of taxon dependent method performance in moths. PLoS One. 10:e0122481.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.
Kisel Y., Mcinnes L., Toomey N.H., Orme C.D.L. 2011. How diversification rates and diversity limits combine to create large-scale species-area relationships. Philos. Trans. R. Soc. B Biol. Sci. 366:2514–2525.

Knowlton N., Weight L.A., Solorzano L.A., Mills D.K., Bermingham E. 1992. Divergence of proteins, mitochondrial DNA and reproductive compatibility across the Isthmus of Panama. Science (80-.). 260:1629–1632.
Kottelat M. 2012. *Rasbora rheophila*, a new species of fish from northern Borneo (Teleostei: Cyprinidae). Rev. suisse Zool. 119:77–87.
Kottelat M. 2013. The fishes of the inland waters of Southeast Asia: A catalogue and core bibliography of the fishes known to occur in freshwaters, mangroves and estuaries. Raffles Bull. Zool.:1–663.
Kottelat M., Britz R., Hui T.H., Witte K.-E. 2006. *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. Proc. R. Soc. B Biol. Sci. 273:895–899.
Kottelat M., Vidthayanon C. 1993. *Boraras micros*, a new genus and species of minute freshwater fish from Thailand (Teleostei: Cyprinidae). Ichthyol. Explor. Freshwaters. 4:161–176.
Kottelat M., Whitten A.J., Kartikasari S.N., Wirjoatmodjo S. 1993. Freshwater fishes of Western Indonesia and Sulawesi. Singapore: Periplus editions.
Kottelat M., Witte K.-E. 1999. Two new species of *Microrasbora* from Thailand and Myanmar, with two new generic names for small Southeast Asian cyprinid fishes (Teleostei: Cyprinidae). J. South Asian Nat. Hist. 4:49–56.
Kumar S., Stecher G., Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33:1870–1874.
Kusuma W.E., Ratmuangkhwang S., Kumazawa Y. 2016. Molecular phylogeny and historical biogeography of the Indonesian freshwater fish *Rasbora lateristriata* species complex (Actinopterygii: Cyprinidae): Cryptic species and west-to-east divergences. Mol. Phylogenet. Evol. 105:212–223.
Lamoreux J.F., Morrison J.C., Ricketts T.H., Olson D.M., Dinerstein E., McKnight M.W., Shugart H.H. 2006. Global tests of biodiversity concordance and the importance of endemism. Nature. 440:212–214.
Landry L., Vincent W.F., Bernatchez L. 2007. Parallel evolution of lake whitefish dwarf ecotypes in association with limnological features of their adaptive landscape. J. Evol. Biol. 20:971–984.
Lavoué Miya, M, Saitoh, K, Ishiguro, NB, Nishida, M S. 2007. Phylogenetic relationships among anchovies, sardines, herrings and their relatives (Clupeiformes), inferred from whole mitogenome sequences. Mol. Phylogenet. Evol. 43:1096–1105.
Lavoué Miya, M., Inoue, J. G., Saitoh, K., Ishiguro, N. B., Nishida, M. S. 2005. Molecular systematics of the Gonorynchiform fishes (Teleostei) based on whole mitogenome sequences: Implications for higher-level relationships within the Otocephala. Mol. Phylogenet. Evol. 37:165–177.
Li F., Li S. 2018. Paleocene-Eocene and Plio-Pleistocene sea-level changes as "species pumps" in Southeast Asia: Evidence from *Althepus* spiders. Mol Phylogenet Evol. 127:545–555.
Liao T.-Y., Kullander S.O., Fang F. 2011. Phylogenetic position of rasborin cyprinids and monophyly of major lineages among the Danioninae, based on morphological characters (Cypriniformes: Cyprinidae). J. Zool. Syst. Evol. Res. 49:224–232.
Liao T.Y., Kullander S.O., Fang F. 2010. Phylogenetic analysis of the genus *Rasbora* (Teleostei: Cyprinidae). Zool. Scr. 39:155–176.
Lim H., Zainal Abidin M., Pulungan C.P., de Bruyn M., Mohd Nor S.A. 2016a. DNA Barcoding Reveals High Cryptic Diversity of the Freshwater Halfbeak Genus *Hemirhamphodon* from Sundaland. PLoS One. 11:e0163596.

Lim N.K.M., Tay Y.C., Srivathsan A., Tan J.W.T., Kwik J.T.B., Baloğlu B., Meier R., Yeo D.C.J. 2016b. Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. R. Soc. Open Sci. 3:160635.
Lisiecki L.E., Raymo M.E. 2005. A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. Paleoceanography. 20:PA1003.
Liu J., Guo X., Chen D., Li J., Yue B., Zeng X. 2019. Diversification and historical demography of the rapid racerunner (*Eremias velox*) in relation to geological history and Pleistocene climatic oscillations in arid Central Asia. Mol Phylogenet Evol. 130:244–258.
Lohman D.J., de Bruyn M., Page T., von Rintelen K., Hall R., Ng P.K.L., Shih H.-T., Carvalho G.R., von Rintelen T. 2011. Biogeography of the Indo-Australian Archipelago. Annu. Rev. Ecol. Evol. Syst. 42:205–226.
Lovejoy N.R., Iranpour M., Collette B.B. 2004. Phylogeny and jaw ontogeny of Beloniform fishes. Integr. Comp. Biol. 44:366–377.
Lumbantobing D.N. 2010. Analisis filogenetik genus *Rasbora* (Teleostei: Cyprinidae) berdasarkan karakter morfologis. J. Iktiollogi Indones. 10:185–189.
Maddison W.P., Maddison D.R. 2019. Mesquite: a modular system for evolutionary analysis. Version 3.61. http://www.mesquiteproject.org.
Maina J.N., Maloiy G.M.O. 1986. The morphology of the respiratory organs of the African air-breathing catfish (*Clarias mossambicus*): A light, electron and scanning microscopic study, with morphometric observations. J. Zool. 209:421–445.
Matzke N.J. 2013. Probabilistic historical biogeography: New models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. Front. Biogeogr. 5:242–248.
Matzke N.J. 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. Syst. Biol. 63:951–970.
Mayden R.L., Tang K.L., Conway K.W., Freyhof J., Chamberlain S., Haskins M., Schneider L., Sudkamp M., Wood R.M., Agnew M., Bufalino A., Sulaiman Z., Miya M., Saitoh K., He S. 2007. Phylogenetic relationships of *Danio* within the order Cypriniformes: a framework for comparative and evolutionary studies of a model species. J. Exp. Zool. 308B:642–654.
McPeek M.A. 2008. The ecological dynamics of clade diversification and community assembly. Am. Nat. 172:e270–e284.
Meisner A.D. 2001. Phylogenetic systematics of the viviparous halfbeak genera *Dermogenys* and *Nomorhamphus* (Teleostei: Hemiramphidae: Zenarchopterinae). Zool. J. Linn. Soc. 133:199–283.
Meyer C.P., Paulay G. 2005. DNA barcoding: Error rates based on comprehensive sampling. PLoS Biol. 3:e422.
Meyer M., Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 2010:pdb-prot5448.
Miller K.G., Miller K.G., Kominz M.A., Browning J. V, Wright J.D., Mountain G.S., Katz M.E., Sugarman P.J., Cramer B.S., Christie-blick N., Pekar S.F. 2005. The Phanerozoic record of global sea-level change. Science (80-.). 310:1293–1298.
Miller M.A., Pfeiffer W., Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 Gatew. Comput. Environ. Work. GCE 2010.
Mittelbach G.G., Schemske D.W., Cornell H. V., Allen A.P., Brown J.M., Bush M.B., Harrison S.P., Hurlbert A.H., Knowlton N., Lessios H.A., McCain C.M., McCune

A.R., McDade L.A., McPeek M.A., Near T.J., Price T.D., Ricklefs R.E., Roy K., Sax D.F., Schluter D., Sobel J.M., Turelli M. 2007. Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. Ecol. Lett. 10:315–31.
Mittermeier R.A., Turner W.R., Larsen F.W., Brooks T.M., Gascon C. 2011. Global biodiversity conservation: The critical role of hotspots. In: Zachos F.E., Habel J.C., editors. Biodiversity hotspots: Distribution and protection of conservation priority areas. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 3–22.
Moreau C.S., Bell C.D. 2013. Testing the museum versus cradle tropical biological diversity hypothesis: Phylogeny, diversification, and ancestral biogeographic range evolution of the ants. Evolution (N. Y). 67:2240–2257.
Moritz C. 1994. Defining 'Evolutionarily Significant Units' for conservation. Trends Ecol. Evol. 9:373–375.
Morlon H., Lewitus E., Condamine F.L., Manceau M., Clavel J., Drury J. 2016. RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. Methods Ecol. Evol. 7:589–597.
Morlon H., Parsons T.L., Plotkin J.B. 2011. Reconciling molecular phylogenies with the fossil record. Proc. Natl. Acad. Sci. 108:16327–16332.
Muchlisin Z.A., Fadli N., Siti-Azizah M.N. 2012. Genetic variation and taxonomy of *Rasbora* group (Cyprinidae) from Lake Laut Tawar, Indonesia. J. Ichthyol. 52:284–290.
Munshi J.S.D. 1961. The accessory respiratory organs of *Clarias batrachus* (Linn.). J. Morphol. 109:115–139.
Myers N., Mittermeier R.A., Mittermeier C.G., da Fonseca G.A.B., Kent J. 2000. Biodiversity hotspots for conservation priorities. Nature. 403:853–858.
Nagel L., Schluter D. 1998. Body size, natural selection, and speciation in sticklebacks. Evolution (N. Y). 52:209–218.
Ng H.H., Kottelat M. 2013. The identity of the cyprinid fishes *Rasbora dusonensis* and *R. tornieri* (Teleostei: Cyprinidae). Zootaxa. 3635:62–70.
Ng H.H., Kottelat M. 2016. The *Glyptothorax* of Sundaland: A revisionary study (Teleostei: Sisoridae). Zootaxa. 4188.
Ng P.K.L., Tay J.B., Lim K.K.P. 1994. Diversity and conservation of blackwater fishes in Peninsular Malaysia, particularly in the North Selangor peat swamp forest. Hydrobiologia. 285:203–218.
Nguyen L.-T., Schmidt H.A., Von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.
Nguyen T.T.T., Na-Nakorn U., Sukmanomon S., ZiMing C. 2008. A study on phylogeny and biogeography of mahseer species (Pisces: Cyprinidae) using sequences of three mitochondrial DNA gene regions. Mol. Phylogenet. Evol. 48:1223–1231.
Nolte A.W., Freyhof J., Stemshorn K.C., Tautz D. 2005. An invasive lineage of sculpins, *Cottus sp.* (Pisces, Teleostei) in the Rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. Proc. R. Soc. B Biol. Sci. 272:2379–2387.
Nores M. 1999. An alternative hypothesis for the origin of Amazonian bird diversity. J. Biogeogr. 26:475–485.
Nores M. 2004. The implications of tertiary and quaternary sea level rise events for avian distribution patterns in the lowlands of northern south america. Glob. Ecol. Biogeogr. 13:149–161.
Nurul Farhana S., Muchlisin Z.A., Duong T.Y., Tanyaros S., Page L.M., Zhao Y., Adamson E.A.S., Khaironizam M.Z., de Bruyn M., Siti Azizah M.N. 2018. Exploring

hidden diversity in Southeast Asia's *Dermogenys spp.* (Beloniformes: Zenarchopteridae) through DNA barcoding. Sci. Rep. 8:10787.
O'Connell K.A., Smart U., Sidik I., Riyanto A., Kurniawan N., Smith E.N. 2019. Diversification of bent-toed geckos (*Cyrtodactylus*) on Sumatra and West Java. Mol. Phylogenet. Evol. 134:1–11.
Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34:2101–2114.
Okonechnikov K., Golosova O., Fursov M., Varlamov A., Vaskin Y., Efremov I., German Grehov O.G., Kandrov D., Rasputin K., Syabro M., Tleukenov T. 2012. Unipro UGENE: A unified bioinformatics toolkit. Bioinformatics. 28:1166–1167.
Ortí G., Meyer A. 1997. The radiation of Characiform fishes and the limits of resolution of mitochondrial ribosomal DNA sequences. Syst. Biol. 46:75–100.
Papadopoulou A., Knowles L.L. 2015a. Genomic tests of the species-pump hypothesis: Recent island connectivity cycles drive population divergence but not speciation in Caribbean crickets across the Virgin Islands. Evolution (N. Y). 69:1501–1517.
Papadopoulou A., Knowles L.L. 2015b. Species-specific responses to island connectivity cycles: Refined models for testing phylogeographic concordance across a Mediterranean Pleistocene Aggregate Island Complex. Mol. Ecol. 24:4252–4268.
Paradis E. 2012. Analysis of phylogenetics and evolution with R. New York: Springer.
Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 20:289–290.
Paradis E., Schliep K. 2019. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.
Patel S., Weckstein J.D., Patane J.S., Bates J.M., Aleixo A. 2011. Temporal and spatial diversification of *Pteroglossus aracaris* (Aves: Ramphastidae) in the neotropics: constant rate of diversification does not support an increase in radiation during the Pleistocene. Mol Phylogenet Evol. 58:105–115.
Pellissier L., Leprieur F., Parravicini V., Cowman P.F., Kulbicki M., Litsios G., Olsen S.M., Wisz M.S., Bellwood D.R., Mouillot D. 2014. Quaternary coral reef refugia preserved fish diversity. Science (80-.). 344:1015–1019.
Pereira L.H.G., Hanner R., Foresti F., Oliveira C. 2013. Can DNA barcoding accurately discriminate megadiverse Neotropical freshwater fish fauna? BMC Genet. 14.
Phillimore A.B., Price T.D. 2008. Density-dependent cladogenesis in birds. PLoS Biol. 6:e71.
Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.
Pouyaud L., Sudarto, Paradis E. 2009. The phylogenetic structure of habitat shift and morphological convergence in Asian *Clarias* (Teleostei, Siluriformes: Clariidae). J. Zool. Syst. Evol. Res. 47:344–356.
Puillandre N., Lambert A., Brouillet S., Achaz G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Mol. Ecol. 21:1864–1877.
Rabosky D.L., Lovette I.J. 2008. Explosive evolutionary radiations: Decreasing speciation or increasing extinction through time? Evolution (N. Y). 62:1866–1875.
Rabosky D.L., Santini F., Eastman J., Smith S.A., Sidlauskas B., Chang J., Alfaro M.E. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nat. Commun. 4:1958.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67:901–904.
Ratnasingham S., Hebert P.D.N. 2007. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Mol. Ecol. Notes. 7:355–364.
Ratnasingham S., Hebert P.D.N. 2013. A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. PLoS One. 8:e66213.
Read C.I., Bellwood D.R., Van Herwerden L. 2006. Ancient origins of Indo-Pacific coral reef fish biodiversity: A case study of the leopard wrasses (Labridae: *Macropharyngodon*). Mol. Phylogenet. Evol. 38:808–819.
Ree R.H., Sanmartín I. 2018. Conceptual and statistical problems with the DEC+J model of founder-event speciation and its comparison with DEC via model selection. J. Biogeogr. 45:741–749.
Ruane S., Bryson R.W., Pyron R.A., Burbrink F.T. 2014. Coalescent species delimitation in milksnakes (genus *Lampropeltis*) and impacts on phylogenetic comparative analyses. Syst. Biol. 63:231–250.
Rüber L., Kottelat M., Tan H., Ng P., Britz R. 2007. Evolution of miniaturization and the phylogenetic position of *Paedocypris*, comprising the world's smallest vertebrate. BMC Evol. Biol. 7.
Rüber L., Tan H.H., Britz R. 2020. Snakehead (Teleostei: Channidae) diversity and the Eastern Himalaya biodiversity hotspot. J. Zool. Syst. Evol. Res. 58:356–386.
Russell E.S. 1931. Some theoretical Considerations on the "Overfishing" Problem. ICES J. Mar. Sci. 6:3–20.
Saitoh Miya, M., Inoue, J. G., Ishiguro, N. B., Nishida, M. K. 2003. Mitochondrial genomics of ostariophysan fishes: perspectives on phylogeny and biogeography. J. Mol. Evol. 56:464–472.
Saitoh K., Sado T., Mayden R., Hanzawa N., Nakamura K., Nishida M., Miya M. 2006. Mitogenomic evolution and interrelationships of the Cypriniformes (Actinopterygii: Ostariophysi): the first evidence toward resolution of higher level relationships of the world's largest freshwater fish clade based on 59 whole mitogenome sequences. J. Mol. Evol. 63:826–841.
Sarr A.C., Sepulchre P., Husson L. 2019. Impact of the Sunda Shelf on the climate of the maritime continent. J. Geophys. Res. Atmos. 124:2574–2588.
Sathiamurthy E., Voris K.H. 2006. Maps of Holocene sea level transgression and submerged lakes on the Sunda Shelf. .
Schipper J., Chanson J.S., Chiozza F., Cox N.A., Hoffmann M., Katariya V., Lamoreux J., Rodrigues A.S.L., Stuart S.N., Temple H.J., Baillie J., Boitani L., Lacher T.E., Mittermeier R.A., Smith A.T., Absolon D., Aguiar J.M., Amori G., Bakkour N., Baldi R., Berridge R.J., Bielby J., Black P.A., Blanc J.J., Brooks T.M., Burton J.A., Butynski T.M., Catullo G., Chapman R., Cokeliss Z., Collen B., Conroy J., Cooke J.G., da Fonseca G.A.B., Derocher A.E., Dublin H.T., Duckworth J.W., Emmons L., Emslie R.H., Festa-Bianchet M., Foster M., Foster S., Garshelis D.L., Gates C., Gimenez-Dixon M., Gonzalez S., Gonzalez-Maya J.F., Good T.C., Hammerson G., Hammond P.S., Happold D., Happold M., Hare J., Harris R.B., Hawkins C.E., Haywood M., Heaney L.R., Hedges S., Helgen K.M., Hilton-Taylor C., Hussain S.A., Ishii N., Jefferson T.A., Jenkins R.K.B., Johnston C.H., Keith M., Kingdon J., Knox D.H., Kovacs K.M., Langhammer P., Leus K., Lewison R., Lichtenstein G., Lowry L.F., Macavoy Z., Mace G.M., Mallon D.P., Masi M., McKnight M.W., Medellin R.A., Medici P., Mills G., Moehlman P.D., Molur S., Mora A., Nowell K., Oates J.F., Olech W., Oliver W.R.L., Oprea M., Patterson B.D.,

Perrin W.F., Polidoro B.A., Pollock C., Powel A., Protas Y., Racey P., Ragle J., Ramani P., Rathbun G., Reeves R.R., Reilly S.B., Reynolds J.E., Rondinini C., Rosell-Ambal R.G., Rulli M., Rylands A.B., Savini S., Schank C.J., Sechrest W., Self-Sullivan C., Shoemaker A., Sillero-Zubiri C., De Silva N., Smith D.E., Srinivasulu C., Stephenson P.J., van Strien N., Talukdar B.K., Taylor B.L., Timmins R., Tirira D.G., Tognelli M.F., Tsytsulina K., Veiga L.M., Vie J.-C., Williamson E.A., Wyatt S.A., Xie Y., Young B.E. 2008. The status of the world's land and marine mammals: Diversity, threat, and knowledge. Science (80-.). 322:225–230.

Schluter D. 2000. Ecological causes of adaptative radiation. Am. Nat. 148:40–64.

Seehausen O. 2007. Evolution and ecological theory: Chance, historical contingency and ecological determinism jointly determine the rate of adaptive radiation. Heredity (Edinb). 99:361–363.

Seehausen O., Takimoto G., Roy D., Jokela J. 2008. Speciation reversal and biodiversity dynamics with hybridization in changing environments. Mol. Ecol. 17:30–44.

Serrao N.R., Steinke D., Hanner R.H. 2014. Calibrating snakehead diversity with DNA barcodes: Expanding taxonomic coverage to enable identification of potential and established invasive species. PLoS One. 9:e99546.

Shen Y., Hubert N., Huang Y., Wang X., Gan X., Peng Z., He S. 2019. DNA barcoding the ichthyofauna of the Yangtze River: Insights from the molecular inventory of a mega-diverse temperate fauna. Mol. Ecol. Resour. 19:1278–1291.

Sholihah A., Delrieu-Trottin E., Sukmono T., Dahruddin H., Risdawati R., Elvyra R., Wibowo A., Kustiati K., Busson F., Sauri S., Nurhaman U., Dounias E., Zein M.S.A., Fitriana Y., Utama I.V., Muchlisin Z.A., Agnèse J.-F., Hanner R., Wowor D., Steinke D., Keith P., Rüber L., Hubert N. 2020. Disentangling the taxonomy of the subfamily Rasborinae (Cypriniformes, Danionidae) in Sundaland using DNA barcodes. Sci. Rep. 10:2818.

Siebert D.J. 1997. The identities of *Rasbora paucisqualis* Ahl in Schreitmuller, 1935, and *Rasbora bankanensis* (Bleeker, 1853), with the designation of a lectotype for R. paucisqualis (Teleostei: Cyprinidae). Raffles Bull. Zool. 45.

Slik J.W.F., Aiba S.I., Bastian M., Brearley F.Q., Cannon C.H., Eichhorn K.A.O., Fredriksson G., Kartawinata K., Laumonier Y., Mansor A., Marjokorpi A., Meijaard E., Morley R.J., Nagamasu H., Nilus R., Nurtjahya E., Payne J., Permana A., Poulsen A.D., Raes N., Riswan S., Van Schaik C.P., Sheil D., Sidiyasa K., Suzuki E., Van Valkenburg J.L.C.H., Webb C.O., Wich S., Yoneda T., Zakaria R., Zweifel N. 2011. Soils on exposed Sunda Shelf shaped biogeographic patterns in the equatorial forests of Southeast Asia. Proc. Natl. Acad. Sci. U. S. A. 108:12343–12347.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30.

Stebbins G.L. 1974. Flowering plants: Evolution above the species level. Harvard University Press.

Stout C.C., Tan M., Lemmon A.R., Lemmon E.M., Armbruster J.W. 2016. Resolving Cypriniformes relationships using an anchored enrichment approach. BMC Evol. Biol. 16:1–13.

Straub S.C.K., Parks M., Weitemier K., Fishbein M., Cronn R.C., Liston A. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am. J. Bot. 99:349–364.

Sullivan J.P., Lavoué S., Hopkins C.D. 2002. Discovery and phylogenetic analysis of

a riverine species flock of African electric fishes (Mormyridae: Teleostei). Evolution (N. Y). 56:597–616.
Swofford D.L. 2001. PAUP*. Phylogenetic Anal. Using Parsimony (*Other Methods). .
Tan H.H. 2020. *Trigonostigma truncata*, a new species of harlequin rasbora from Malay Peninsula (Teleostei: Danionidae). RAFFLES Bull. Zool. 7600.
Tan H.H., Lim K.K.P. 2013. Three new species of freshwater halfbeaks (Teleostei: Zenarchopteridae: *Hemirhamphodon*) from Borneo. Raffles Bull. Zool. 61:735–747.
Tan M., Armbruster J.W. 2018. Phylogenetic classification of extant genera of fishes of the order Cypriniformes (Teleostei: Ostariophysi). Zootaxa. 4476:6.
Tan M.P., Jamsari A.F.J., Siti Azizah M.N. 2012. Phylogeographic pattern of the striped snakehead, *Channa striata* in Sundaland: Ancient river connectivity, geographical and anthropogenic signatures. PLoS One. 7:1–11.
Tang K.L., Agnew M.K., Hirt M.V., Sado T., Schneider L.M., Freyhof J., Sulaiman Z., Swartz E., Vidthayanon C., Miya M., Saitoh K., Simons A.M., Wood R.M., Mayden R.L. 2010. Systematics of the subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). Mol. Phylogenet. Evol. 57:189–214.
den Tex R.J., Leonard J.A. 2013. A molecular phylogeny of Asian barbets: Speciation and extinction in the tropics. Mol. Phylogenet. Evol. 68:1–13.
den Tex R.J., Thorington R., Maldonado J.E., Leonard J.A. 2010. Speciation dynamics in the SE Asian tropics: Putting a time perspective on the phylogeny and biogeography of Sundaland tree squirrels, Sundasciurus. Mol. Phylogenet. Evol. 55:711–720.
Tilak M.-K., Justy F., Debiais-Thibaud M., Botero-Castro F., Delsuc F., Douzery E.J.P. 2015. A cost-effective straightforward protocol for shotgun Illumina libraries designed to assemble complete mitogenomes from non-model species. Conserv. Genet. Resour. 7:37–40.
Verheyen E., Salzburger W., Snoeks J., Meyer A. 2003. Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science (80-.). 300:325–329.
Vogler A.P., Desalle R. 1994. Diagnosing units of conservation management. Conserv. Biol. 8:354–363.
Voris H.K. 2000. Maps of Pleistocene sea levels in Southeast Asia: Shorelines, river systems and time durations. J. Biogeogr. 27:1153–1167.
Walker J.D., Geissman J.W., Bowring S.A., Babcock L.E. 2018. GSA Geologic Time Scale v. 5.0. .
Wallace A.R. 1869. The Malay Archipelago: The land of the orang-utan and the bird of paradise, a narrative of travel, with studies of man and nature. London: Harper.
Weigelt P., Steinbauer M.J., Cabral J.S., Kreft H. 2016. Late Quaternary climate change shapes island biodiversity. Nature. 532:99–102.
Weir J.T., Schluter D. 2007. The latitudinal gradient in recent speciation and extinction rates of birds and mammals. Science (80-.). 315:1574–1576.
Westerhold T., Marwan N., Drury A.J., Liebrand D., Agnini C., Anagnostou E., Barnet J.S.K., Bohaty S.M., De Vleeschouwer D., Florindo F. 2020. An astronomically dated record of Earth's climate and its predictability over the last 66 million years. Science (80-.). 369:1383–1387.
Wiens J.J., Donoghue M.J. 2004. Historical biogeography, ecology and species richness. Trends Ecol. Evol. 19:639–644.
Woodruff D.S. 2010. Biogeography and conservation in Southeast Asia: How 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugial-phase biodiversity. Biodivers. Conserv.

19:919–941.
Wurster C.M., Rifai H., Zhou B., Haig J., Bird M.I. 2019. Savanna in equatorial Borneo during the late Pleistocene. Sci. Rep. 9:1–7.
Zachos F.E. 2018. Species concepts and species delimitation in mammals. In: Zachos F.E., Asher R.J., editors. Mammalian evolution, diversity and systematics. Berlin: de Gruyter. p. 1–13.
Zachos J.C., Dickens G.R., Zeebe R.E. 2008. An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. Nature. 451:279–283.
Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics. 29:2869–2876.

Appendix A
Revisiting species boundaries and distribution ranges of *Nemacheilus spp.* (Cypriniformes: Nemacheilidae) and *Rasbora spp.* (Cypriniformes: Cyprinidae) in Java, Bali and Lombok through DNA barcodes: implications for conservation in a biodiversity hotspot

Nicolas Hubert[1], Daniel Lumbantobing[2], Arni Sholihah[1,3], Hadi Dahruddin[1,2], Erwan Delrieu-Trottin[1,4], Frédéric Busson[1,5], Sopian Sauri[2], Renny Hadiaty[2], Philippe Keith[5]*

1 Institut de Recherche pour le Développement, UMR 226 ISEM (UM, CNRS, IRD, EPHE), Université de Montpellier, Place Eugène Bataillon, CC 065, 34095 Montpellier cedex 05, France
2 Division of Zoology, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Jalan Raya Jakarta Bogor Km 46, Cibinong 16911, Indonesia
3 Institut Teknologi Bandung, School of Life Sciences and Technology, Bandung, Indonesia
4 Museum für Naturkunde, Leibniz-Institut für Evolutions und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Invalidenstrasse 43, 10115 Berlin, Germany
5 UMR 7208 BOREA (MNHN-CNRS-UPMC-IRD-UCBN), Muséum National d'Histoire Naturelle, 43 rue Cuvier, 75231 Paris cedex 05, France
*email: nicolas.hubert@ird.fr

Conservation Genetics volume 20 issue 3 pages 517-529 (2019)
https://doi.org/10.1007/s10592-019-01152-w
Received: 27 September 2018; Accepted: 31 January 2019;
Published online: 13 February 2019

Conservation Genetics
https://doi.org/10.1007/s10592-019-01152-w

RESEARCH ARTICLE

Revisiting species boundaries and distribution ranges of *Nemacheilus* spp. (Cypriniformes: Nemacheilidae) and *Rasbora* spp. (Cypriniformes: Cyprinidae) in Java, Bali and Lombok through DNA barcodes: implications for conservation in a biodiversity hotspot

Nicolas Hubert[1] · Daniel Lumbantobing[2] · Arni Sholihah[1,3] · Hadi Dahruddin[1,2] · Erwan Delrieu-Trottin[1,4] · Frédéric Busson[1,5] · Sopian Sauri[2] · Renny Hadiaty[2] · Philippe Keith[5]

Received: 27 September 2018 / Accepted: 31 January 2019
© Springer Nature B.V. 2019

Abstract

Biodiversity hotspots have provided useful geographic proxies for conservation efforts. Delineated from a few groups of animals and plants, biodiversity hotspots do not reflect the conservation status of freshwater fishes. With hundreds of new species described on a yearly basis, fishes constitute the most poorly known group of vertebrates. This situation urges for an acceleration of the fish species inventory through fast and reliable molecular tools such as DNA barcoding. The present study focuses on the freshwater fishes diversity in the Sundaland biodiversity hotspot in Southeast Asia. Recent studies evidenced large taxonomic gaps as well as unexpectedly high levels of cryptic diversity, particularly so in the islands of Java and Bali. The Cypriniformes genera *Rasbora* and *Nemacheilus* account for most of the endemic species in Java and Bali, however their taxonomy is plagued by confusion about species identity and distribution. This study examines the taxonomic status of the *Rasbora* and *Nemacheilus* species in Java, Bali and Lombok islands through DNA barcodes, with the objective to resolve taxonomic confusion and identify trends in genetic diversity that can be further used for conservation matters. Several species delimitation methods based on DNA sequences were used and confirmed the status of most species, however several cases of taxonomic confusion and two new taxa are detected. Mitochondrial sequences argue that most species range distributions currently reported in the literature are inflated due to erroneous population assignments to the species level, and further highlight the sensitive conservation status of most *Rasbora* and *Nemacheilus* species on the islands of Java, Bali and Lombok.

Keywords Conservation genetics · Taxonomy · Southeast Asia · Cryptic diversity · Population fragmentation

Introduction

Biodiversity hotspots are characterized by high proportions of endemic species and high levels of anthropogenic threats (Myers et al. 2000). Identified to maximize conservation efforts in a world with finite human and funding

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10592-019-01152-w) contains supplementary material, which is available to authorized users.

✉ Nicolas Hubert
nicolas.hubert@ird.fr

[1] Institut de Recherche pour le Développement, UMR 226 ISEM (UM, CNRS, IRD, EPHE), Université de Montpellier, Place Eugène Bataillon, CC 065, 34095 Montpellier cedex 05, France

[2] Division of Zoology, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Jalan Raya Jakarta Bogor Km 46, Cibinong 16911, Indonesia

[3] Instut Teknologi Bandung, School of Life Science and Technology, Bandung, Indonesia

[4] Museum für Naturkunde, Leibniz-Institut für Evolutions und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Invalidenstrasse 43, 10115 Berlin, Germany

[5] UMR 7208 BOREA (MNHN-CNRS-UPMC-IRD-UCBN), Muséum National d'Histoire Naturelle, 43 rue Cuvier, 75231 Paris cedex 05, France

Published online: 13 February 2019

resources for conservation matters, biodiversity hotspots have provided useful geographic proxies for conservation efforts. While those biodiversity hotspots have been delineated based on a limited set of well-known vertebrate taxa such as mammals, birds, amphibians and reptiles, the diversity and status of the world's most diverse vertebrate group, that is fishes, is still largely unknown (Myers et al. 2000; Lamoreux et al. 2006; Hoffman et al. 2010). With hundreds of new species described on a yearly basis, freshwater fishes suffer from an important taxonomic knowledge gaps that, combined with the taxonomic impediment (i.e. the rarefaction of taxonomists worldwide), currently plagues conservation efforts in most biodiversity hotspots (Winemiller et al. 2016; Garnett and Christidis 2017). This situation arguably accounts for their exclusion from most of the large-scale meta-analyses conducted so far on global diversity patterns (Myers et al. 2000; Lamoreux et al. 2006; Hoffman et al. 2010).

In insular South-East Asia (SEA), the Sundaland hotspot exemplifies the stakes faced by conservation stakeholders due to antagonistic interests in the use of biological resources. Including the islands of Java, Sumatra and Borneo, Sundaland is currently among the largest hotspots in terms of number of species and endemics (Myers et al. 2000). Recent threat analyses, however rank it as one of the most threatened (Lamoreux et al. 2006; Hoffman et al. 2010). With nearly 900 species and 430 endemics, Sundaland accounts respectively for 74% and 48% of the total and endemic diversity of the approximately 1200 fish species cited from rivers of the Indonesian archipelago (Hubert et al. 2015). Within Sundaland, Java exhibits one the highest fish species density with 1.7 species/1000 km^2 (213 species) together with Sumatra (460 species) and ahead of Kalimantan (Indonesian Borneo; 1.2 species per 1000 km^2 and 646 species). Hosting 130 million of people sharing 130,000 km^2, Javanese aquatic ecosystems have faced a dramatic increase of anthropogenic threats during the last decade. The recent molecular inventory of the Javanese ichthyofauna evidenced large discrepancies between the checklist of Java freshwater fishes established from historical records (Hubert et al. 2015) and a modern reappraisal based on DNA sequences (Dahruddin et al. 2017), hence highlighting major gaps in the taxonomic knowledge of this ichthyofauna. Along the same line Hutama et al. (2017) evidenced high levels of cryptic diversity (i.e. morphologically unnoticed diversity) in widespread fish species of Java deriving from a late Pleistocene fragmentation of the populations associated with population bottlenecks. Considering that Sundaland is currently in a refugial state and that its emerged lands represent only a small fraction of its average surface during the Pleistocene (Woodruff 2010; Lohman et al. 2011), the state of Sundaland ichthyofauna urges for an acceleration of the ichthyological exploration of its freshwaters.

Initially designed to circumvent the taxonomic impediment by proposing a standard molecular framework for species identification through the use of the mitochondrial cytochrome oxidase I gene as an internal species tag, DNA barcoding opened new perspectives in the inventory of freshwater fishes (Hubert et al. 2008; Ward et al. 2009; Steinke and Hanner 2011). While large scale fish DNA barcoding campaigns have been tackled during the last decade (April et al. 2011; Hubert et al. 2012, 2018; Pereira et al. 2013; Geiger et al. 2014; Knebelsberger et al. 2015; Dahruddin et al. 2017; Durand et al. 2017; Machado et al. 2018), it becomes more and more evident that the pace of species description is surpassed by the astonishing underestimation of species diversity, often referring to cryptic diversity, and the complexity of fish biodiversity (Hubert et al. 2012; Jaafar et al. 2012; Kadarusman et al. 2012; Geiger et al. 2014; Winterbottom et al. 2014). We focus in the present study on the diversity and range distribution in South Sundaland of two Cypriniformes genera, namely *Rasbora* (Cyprinidae) and *Nemacheilus* (Nemacheilidae) that constitute emblematic endemic lineages in Java and Lesser Sunda Islands (Bali, Lombok) due to their occurrence in a large array of aquatic ecosystems and their high levels of endemism compared to other genera occurring in Java. Mostly described during the eighteenth and nineteenth centuries, *Rasbora* and *Nemacheilus* taxonomy and distribution is confusing in Java due to the lack of traceability of the taxonomic information often associated with old descriptions. Type localities are available for most of these species (Kottelat 2013), however range distribution are currently unknown (Froese and Pauly 2014; Hubert et al. 2015; Eschmeyer et al. 2018), most *Rasbora* and *Nemacheilus* species being reported in Java and/or Bali without further details. With the aim to re-examine *Rasbora* and *Nemacheilus* diversity on the islands of Java, Bali and Lombok, we produced a DNA barcode reference library with the following objectives: (1) exploration of species biological boundaries through DNA-based species delimitation methods, (2) validation of species identity and taxonomy and precise range distribution by producing DNA barcodes from type localities or neighboring watersheds, (3) estimation of species genetic diversity and production of recommendations for conservation genetics purposes.

Materials and methods

Sampling and collection management

The authors previously conducted a large-scale DNA barcoding campaign across 95 sites in Java and Bali Island between November 2012 and May 2015 (Dahruddin et al. 2017). During this initial inventory, a total of 3310 specimens, including 162 species belonging to 110 genera and 53

families were collected. This was complemented by an additional campaign in Lombok island on March 2015 resulting in the sampling of an additional set of 367 specimens belonging to 54 species and 44 genera sampled across 12 sites. With the objective to produce a DNA barcode reference library for the Java and Bali ichthyofauna, a total of 24 specimens for 4 species of *Rasbora* and 15 specimens for 2 species of *Nemacheilus* were previously sequenced (Dahruddin et al. 2017). Considering the objectives of the present study, an additional set of 84 specimens of *Nemacheilus* and 118 specimens of *Rasbora* were selected at all the sites these genera were sampled during the initial campaign for further sequencing (Fig. 1). Thus, a total of 99 specimens belonging to 2 species of *Nemacheilus* and 142 specimens belonging to 4 species of *Rasbora* were analyzed in the present study (Table S1).

Specimens were captured using various gears including electrofishing, seine nets, cast nets and gill nets across sites encompassing the diversity of freshwater lentic and lotic habitats. Specimens were identified following available monographs (Kottelat et al. 1993), and species names were further validated based on several online catalogues (Froese and Pauly 2014; Eschmeyer et al. 2018). Specimens were photographed and individually labeled, and voucher specimens were preserved in a 5% formalin solution. A fin clip or a muscle biopsy was taken for each specimen and fixed in a 96% ethanol solution for genetic analyses. Both tissues and voucher specimens were deposited in the national collections at the Muzeum Zoologicum Bogoriense (MZB) in the Research Centre for Biology (RCB) from the Indonesian Institute of Sciences (LIPI).

Sequencing and international repositories

Genomic DNA was extracted using a Qiagen DNeasy 96 tissue extraction kit following the manufacturer's specifications. A 651-bp segment from the 5′ region of the cytochrome oxidase I gene (COI) was amplified using primer cocktails C_FishF1t1/C_FishR1t1 including M13 tails (Ivanova et al. 2007). PCR amplifications were done on a Veriti 96-well Fast (ABI-AppliedBiosystems) thermocycler with a final volume of 10.0 µl containing 5.0 µl Buffer 2×, 3.3 µl ultrapure water, 1.0 µl each primer (10 µM), 0.2 µl enzyme Phire® Hot Start II DNA polymerase (5 U) and 0.5 µl of DNA template (~ 50 ng). Amplifications were conducted as follow: initial denaturation at 98 °C for 5 min followed by 30 cycles denaturation at 98 °C for 5 s, annealing at 56 °C for 20 s and extension at 72 °C for 30 s, followed by a final extension step at 72 °C for 5 min. The PCR products were purified with ExoSap-IT® (USB Corporation, Cleveland, OH, USA) and sequenced in both directions. Sequencing reactions were performed using the "BigDye® Terminator v3.1 Cycle Sequencing Ready Reaction" and sequencing was performed on the automatic sequencer ABI

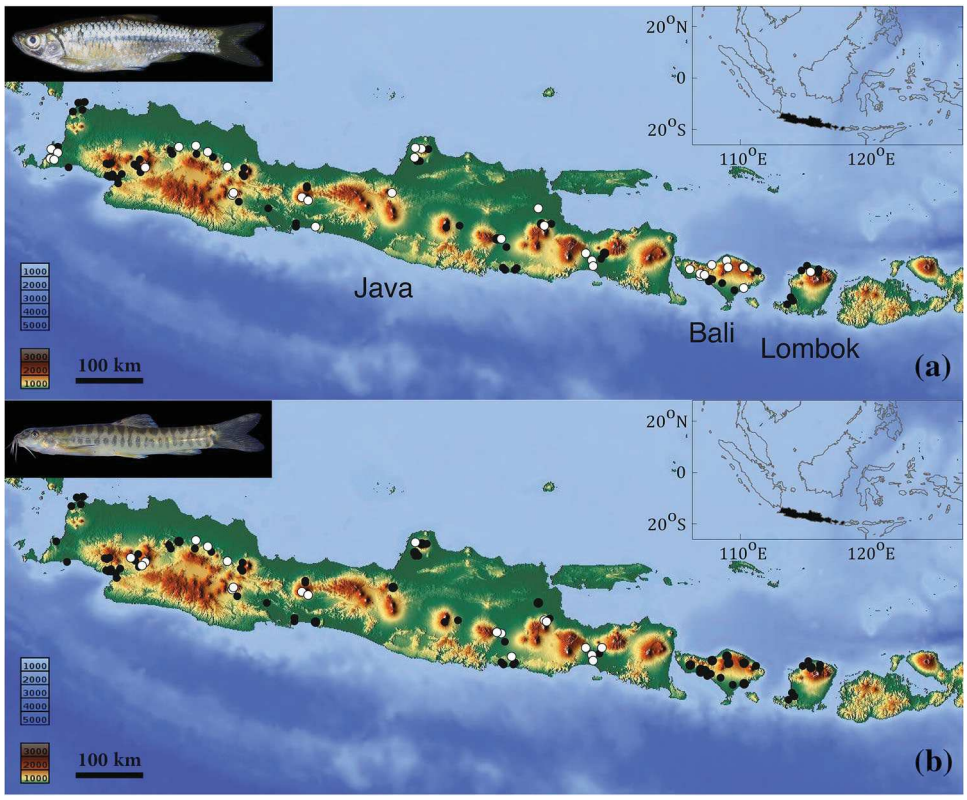

Fig. 1 Collection sites for the 241 samples analyzed in the present study following the sampling campaign detailed in Dahruddin et al. 2017 and new sampling events in Lombok island. **a** Collection sites of *Rasbora* specimens. **b** Collection sites of *Nemacheilus* specimens. White dots correspond to sites where *Rasbora* or *Nemacheilus* specimens were collected. Black dots represent visited sites where no *Rasbora* or *Nemacheilus* specimens were observed. Each dot may represent several collection sites

3130 DNA Analyzer (Applied Biosystems). The sequences and collateral information have been deposited in BOLD (Ratnasingham and Hebert 2007) and are available in the projects BIFH, BIFHB, BIFI and BIFB. DNA sequences were submitted to GenBank (accession numbers are accessible directly at the individual records in BOLD).

Species delimitation and genetic diversity

A maximum likelihood (ML) tree was first reconstructed using phyml 3.0.1 (Guindon and Gascuel 2003) based on the most likely substitution model selected by JMODELTEST 2.1.7 (Darriba et al. 2012). An ultrametric and fully resolved tree was reconstructed using the Bayesian approach implemented in BEAST 2.4.8 (Bouckaert et al. 2014). Two markov chain of 50 million each were ran independently using Yule pure birth model tree prior and an uncorrelated relaxed lognormal clock model for both *Rasbora* and *Nemacheilus* data sets. The ML tree was converted into an ultrametric tree using a relaxed clock model of the chronos function in the R package ape 4.1 (Paradis 2004) implemented in R (R Core Team 2018) and further used to initiate tree searches for the Bayesian analyses. Calibrations of ML and Bayesian analyses were established following Hutama et al. (2017). Age intervals for the Most Recent Common Ancestor (MRCA) of *Rasbora* spp. and *Nemacheilus* spp. were estimated based on the canonical 1.2% (+/− 0.5%) of genetic distance per million years for the fish COI gene (Bermingham et al. 1997). The average genetic distances between species pairs involving a direct ancestry with the MRCAs of *Rasbora* and *Nemacheilus* were calculated using MEGA 6 (Tamura et al. 2013) and used to estimate the age interval of the MRCAs. An additional calibration was added in the *Rasbora* tree including *Rasbora baliensis, R. lateristriata* and *R. aprotaenia* and also in *Nemacheilus* tree for the MRCA of *N. chrysolaimos* haplotypes following the same methodology. Trees were sampled every 10,000 states after an initial burning period of 10 million and both runs were combined using LogCombiner 2.4.8 (Bouckaert et al. 2014). The maximum credibility tree was constructed using TreeAnnotator 2.4.7 (Bouckaert et al. 2014).

Several alternative methods have been proposed for delimitating molecular lineages (Pons et al. 2006; Puillandre et al. 2012; Ratnasingham and Hebert 2013; Zhang et al. 2013; Hubert and Hanner 2015). These methods rely on different approaches and assumptions but they all have in common the detection of transitions between mutation/drift (within species) and speciation/extinction (between species) dynamics (Hubert and Hanner 2015). Each of these methods is prone to pitfalls, particularly regarding singletons (i.e. delimitated lineages represented by a single sequence) and combining different approaches is increasingly used to circumvent potential pitfalls arising from, for instance, uneven sampling among species (Kekkonen and Hebert 2014; Kekkonen et al. 2015; Blair and Bryson 2017). Here, four sequence-based methods of species delimitation were used to delimitate species, and a final delimitation scheme was established based on a 50% consensus among methods in order to produce a robust delimitation scheme. For the sake of clarity, species identified based on morphological characters are referred to as species while species delimitated by DNA sequences are referred to as Operational Taxonomic Units (OTU), defined as diagnosable molecular lineages (Avise 1989; Moritz 1994; Vogler and DeSalle 1994; Hutama et al. 2017). OTUs were delimitated using the following algorithms: (1) Refined single linkage (RESL) as implemented in BOLD and used to produce Barcode Index Numbers (BIN)(Ratnasingham and Hebert 2013), (2) Automatic barcode gap discovery (ABGD) (Puillandre et al. 2012), (3) Poisson tree process (PTP) in its multiple rates version (mPTP) as implemented in the standalone software mptp_0.2.3 (Zhang et al. 2013; Kapli et al. 2017), and (4) General mixed yule-coalescent (GMYC) in its single rate version (sGMYC) as implemented in the R package splits 1.0-19 (Ezard et al. 2009; Fujisawa and Barraclough 2013). RESL and ABGD used the DNA alignments as inputs while the ML tree was used for mPTP. Two delimitation schemes were collected for sGMYC: (1) a scheme based on the maximum credibility tree from the Bayesian analysis as input (sGMYC), (2) a consensus scheme with OTUs selected if present in more than 50% of the 10 replicates of sGMYC based on 10 Bayesian trees sampled along the Markov chain (sGMYC*).

We quantified the match among methods and their relative power using the *match ratio*, the Relative Taxonomic Index of Congruence index (R_{tax}) and the Taxonomic Index of Congruence (C_{tax}) following Blair and Bryson (2017). The match ratio is a measure of concordance among methods and is defined as twice the number of matches divided by the sum of the number of delimitated OTUs and the number of morphological species (Arhens et al. 2016). The R_{tax} index quantifies the relative power of a method to infer all estimated speciation events and is defined as the number of speciation events identified by a method divided by the total number of speciation events identified by the different methods (Miralles and Vences 2013). The C_{tax} index is a measure of congruence in species assignments between two methods and is calculated by dividing the number of speciation events inferred jointly by the two methods by the total number of speciation events inferred. Considering the number of comparisons involved, an average C_{tax} index was calculated for each method.

For each species, Kimura 2-parameter (K2P) pairwise genetic distances were calculated using the R package ape 4.1 (Paradis 2004). Maximum intraspecific and nearest neighbor genetic distances were calculated from the matrice of pairwise K2P genetic distances using the R package

SPIDER 1.5 (Brown et al. 2012). Haplotype diversity (*h*) and nucleotide diversity (π) were calculated for each species using the R package pegas 0.1 (Paradis 2010).

Results

99 and 142 sequences were successfully obtained for *Nemacheilus* and *Rasbora* respectively. All the sequences were above 500 bp of length and no stop codons were detected, suggesting that the sequences collected represent functional coding regions. The maximum credibility tree of *Rasbora* spp. identified a group of closely related species including *R. aprotaenia, R. lateristriata* and *R. baliensis* as well as two unknown taxa labeled as *R*. sp1 and *R*. sp2 (Fig. 2). The age of the MRCA of this clade of closely related species is inferred to trace back around 3 million years ago (Ma), and the split between *Rasbora argyrotaenia* and the remaining *Rasbora* is inferred to happen around 11 Ma. The age of *Rasbora* species MRCAs ranged between

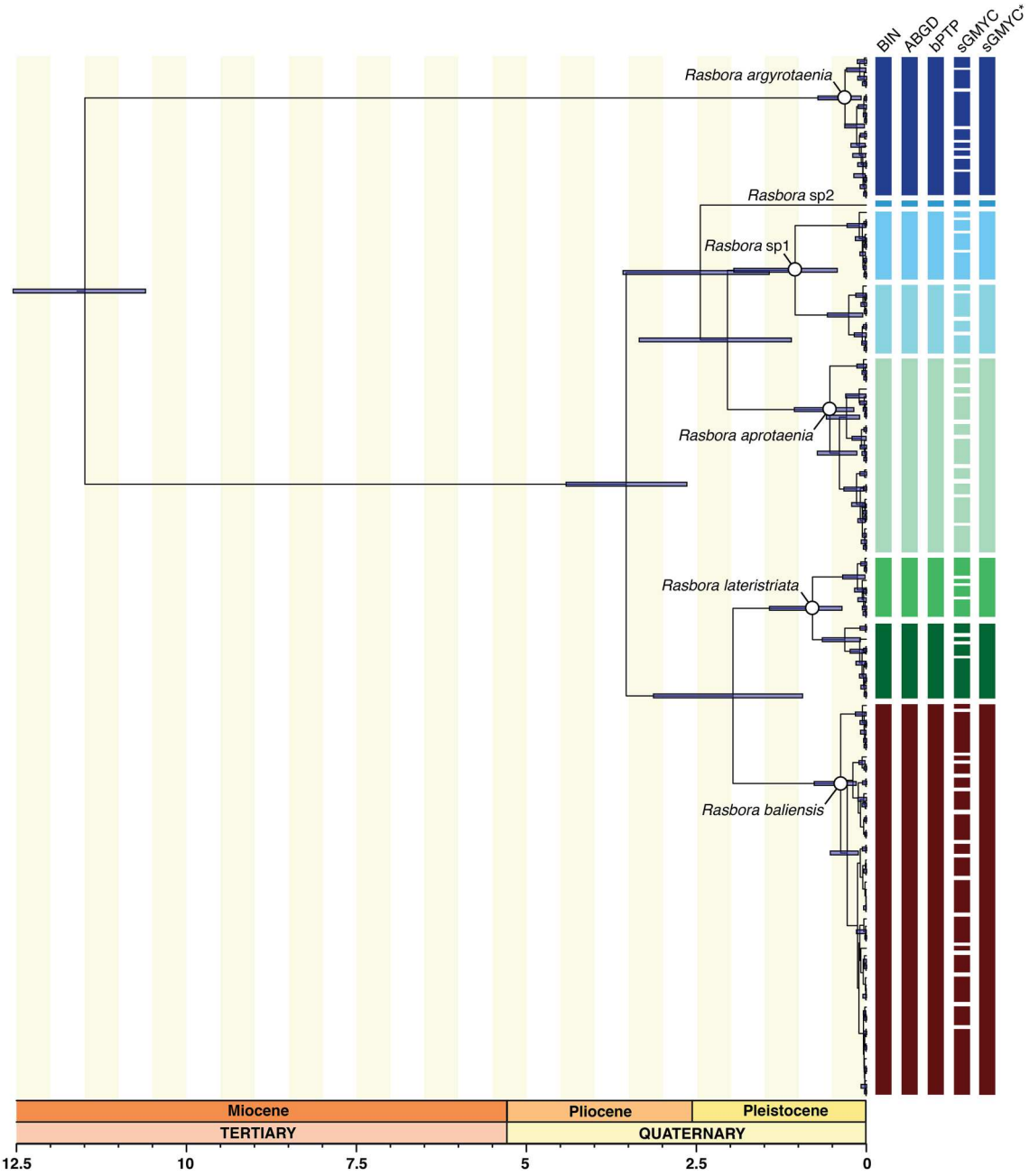

Fig. 2 Bayesian maximum credibility tree of *Rasbora* DNA barcodes including 95% HPD interval for node age estimates and sequence clustering results according to the 5 species delimitation methods implemented

0.5 Ma for *R. baliensis* and 1 Ma for *R.* sp1. The maximum credibility tree clearly separated the two *Nemacheilus* species genealogy into two distinct clades with a MRCA dated around 10 Ma (Fig. 3). The MRCA of *N. chrysolaimos* and *N. fasciatus* genealogies are dated around 1.5 and 0.5 Ma respectively.

Delimitation methods largely converged in identifying 8 OTUs within the 6 species of *Rasbora* recognized here. Of the two partitioning schemes obtained with sGMYC, only the consensus partitioning scheme derived from 10 replicates (sGMYC*) is consistent with other methods in *Rasbora* (Fig. 1), with a number of OTUs ranging from 7 to 9 across sampled trees. Applying sGMYC to the Bayesian maximum credibility tree resulted in an inflated number of OTUs as 51 lineages were delineated (Table 1). Two OTUs were detected within *R. lateristriata* and *R.* sp1. The *match ratio* was similar among methods excepting sGMYC and the highest resolution power was observed for sGMYC with a R_{tax} of 1 (Table 1). The highest taxonomy congruence was observed for BIN, ABGD, mPTP and sGMYC* with a C_{tax} of 0.784 (Table 1). Delimitation methods produced concordant delimitation schemes within *Nemacheilus*, as all methods, excepting sGMYC, delineated one OTU for each of the two species (Fig. 3). As observed for *Rabora*, sGMYC inflated the number of OTUs with 4 OTUs delimitated within *N. chrysolaimos* (Table 2). The *match ratio* and the taxonomic concordance (C_{tax}) were the highest for all methods excepting sGMYC (Table 2). The resolution power was estimated to be the highest for sGMYC as revealed by R_{tax} (Table 2).

Maximum intraspecific K2P distances ranged between 0.62 in *R. argyrotaenia* and 2.51 in *R.* sp1 (Table 1), and the

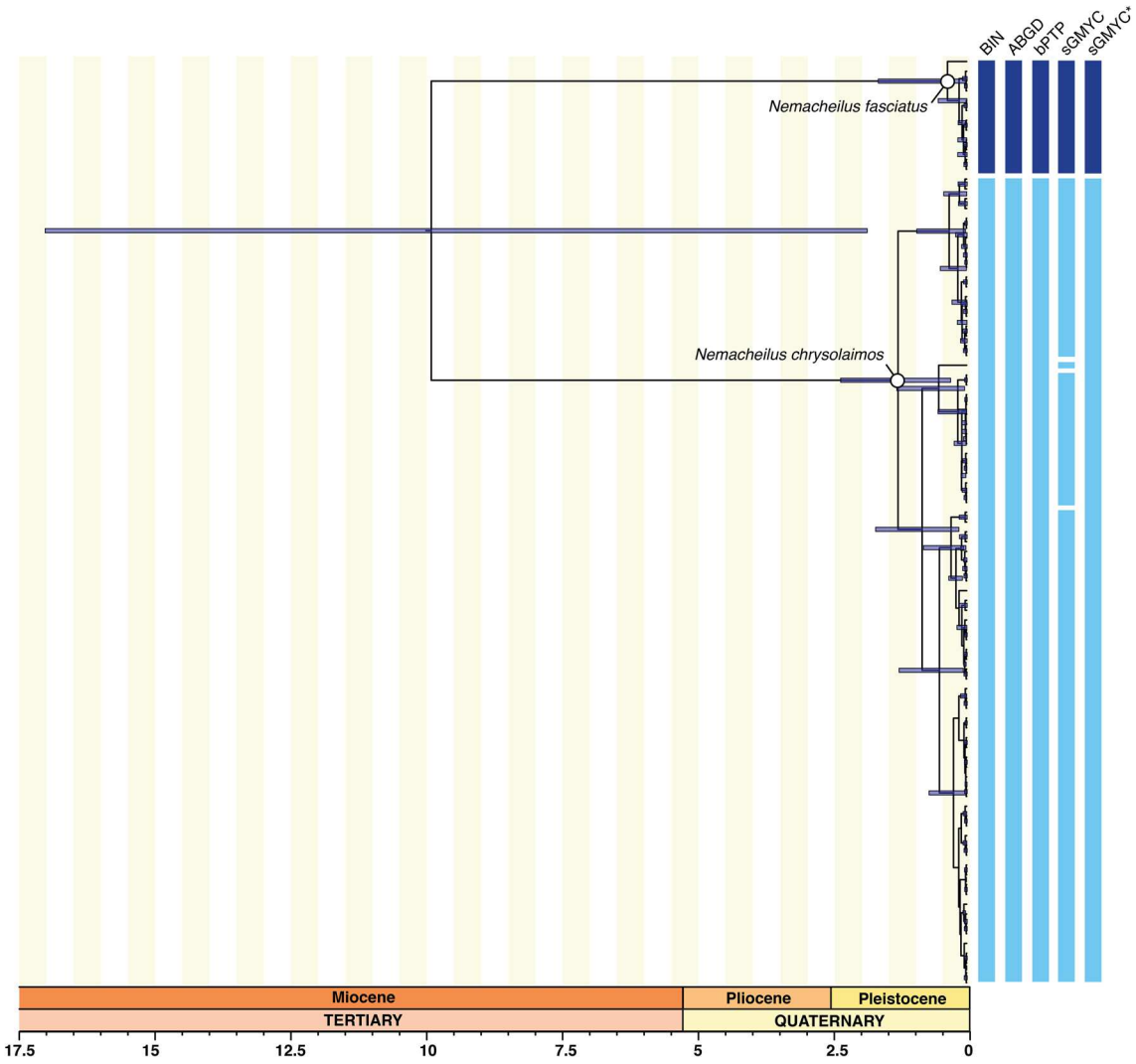

Fig. 3 Bayesian Maximum Credibility Tree of *Nemacheilus* DNA barcodes including 95% HPD interval for node age estimates and sequence clustering according to the 5 species delimitation methods implemented

Table 1 Summary statistics of *Rasbora* species genetic diversity and species delimitation schemes

Taxon	*n*	Max. K2P distance	Min. nearest neighbor K2P distance	*h*	π	BIN	ABGD	bPTP	sGMYC	sGMYC*
R. argyrotaenia	20	0.62	9.67	0.742	0.001	1	1	1	8	1
R. aprotaenia	27	0.93	2.68	0.795	0.004	1	1	1	10	1
R. lateristriata	20	2.08	3.04	0.753	0.009	2	2	2	8	2
R. baliensis	54	0.77	3.04	0.505	0.002	1	1	1	16	1
R. sp1	20	2.51	3.33	0.658	0.012	2	2	2	8	2
R. sp2	1	NA	NA	NA	NA	1	1	1	1	1
Total	142	–	–	–	–	8	8	8	51	8
Match ratio	–	–	–	–	–	0.571	0.571	0.571	0.035	0.571
R_{tax}	–	–	–	–	–	0.16	0.16	0.16	1	0.16
Mean C_{tax}	–	–	–	–	–	0.784	0.784	0.784	0.04	0.784

n, number of individual sequences analyzed; Max. K2P distance, maximum K2P genetic distance (percent); Min. Nearest Neighbor K2P distance (percent), minimum K2P genetic distance to the nearest phylogenetic neighbor (percent) in the present data set; *h*, haplotypic diversity; π, nucleotidic diversity; match ratio, twice the number of match divided by the sum of the number of delimitated OTU and the number of morphological species; R_{tax}, number of speciation events identified by a method divided by the total number of speciation events identified by the different methods; C_{tax} , average number of speciation events inferred jointly by two methods divided by the total number of speciation events inferred

Table 2 Summary statistics of *Nemacheilus* species genetic diversity and species delimitation schemes

Taxon	*n*	Max. K2P distance	Min. nearest neighbour K2P distance	*h*	π	BIN	ABGD	bPTP	sGMYC	sGMYC*
N. fasciatus	14	0.47	14.78	0.168	0.001	1	1	1	1	1
N. chrysolaimos	85	1.72	14.78	0.828	0.007	1	1	1	4	1
Total	99	–	–	–	–	2	2	2	5	2
Match ratio	–	–	–	–	–	1	1	1	0.286	1
R_{tax}	–	–	–	–	–	0.25	0.25	0.25	1	0.25
Mean C_{tax}	–	–	–	–	–	0.813	0.813	0.813	0.25	0.813

n, number of individual sequences analyzed; Max. K2P distance, maximum K2P genetic distance (percent); Min. Nearest Neighbor K2P distance (percent), minimum K2P genetic distance to the nearest phylogenetic neighbor (percent) in the present data set; *h*, haplotypic diversity; π, nucleotidic diversity; match ratio, twice the number of match divided by the sum of the number of delimitated OTU and the number of morphological species; R_{tax}, number of speciation events identified by a method divided by the total number of speciation events identified by the different methods; C_{tax}, average number of speciation events inferred jointly by two methods divided by the total number of speciation events inferred

minimum nearest neighbor K2P distance ranged between 2.68 in *R. aprotaenia* and 9.67 in *R. argyrotaenia*. Haplotype diversity was globally high for all species with a haplotype diversity ranging from 0.795 in *R. aprotaenia* to 0.505 in *R. baliensis*. Nucleotide diversity was low for all species, with the lowest values ranging from 0.001 in *R. argyrotaenia* to 0.004 in *R. aprotaenia*, excepting for *R. lateristriata* and *R.* sp1, the only two species with two OTUs (Table 1). Maximum intraspecific K2P distances were 0.47 and 1.72 for *N. fasciatus* and *N. chrysolaimos*, respectively and the K2P distance between the two species was 14.78 (Table 2). Genetic diversity is markedly different between the two *Nemacheilus* species with *N. fasciatus* exhibiting a low genetic diversity with a haplotype diversity of 0.168 and a nucleotide diversity of 0.001 while *N. chrysolaimos* has a haplotype diversity of 0.828 and a nucleotide diversity of 0.007 (Table 2).

Discussion

Taxonomy of *Rasbora* species in Java, Bali and Lombok

Species boundaries of the four *Rasbora* species were successfully recovered by all, except sGMYC, species delimitation methods. This result has several implications regarding the taxonomy of the *Rasbora* genus in Java. The morphological characters described by Kottelat (1993) were not all operational as different and non-standardized meristic features were described for each species. Coloration patterns were used to propose an initial set of morphological identifications that resulted in the acknowledgement of four species based on the following key: (1) *R. aprotaenia* is distinguished by two dark spots along the lateral line being connected by a thin and diffuse dark line, the first below

the origin of the dorsal fin, the second on the caudal peduncle and a dark spot along the proximal margin of the anal fin (Fig. 4). (2) *R. argyrotaenia* can be further separated in having a continuous dark line that cover nearly entirely the lateral line and a dark line that underlines the entire proximal margin of the anal fin. (3) *R. laterstriata* and *R. baliensis* are further distinguished by the number of scale on the lateral line with 26–28 in *R. baliensis* and 29–33 in *R. lateristriata* (Kottelat et al. 1993). The concordance between the DNA-based and coloration- and meristic-based delimitation schemes for the four known *Rasbora* species confirms their biological species status. In addition, the examination of range distributions and known type localities further confirm this concordance with type localities being contained within the observed distribution range for *R. argyrotaenia, R. aprotenia* and *R. baliensis* (Fig. 5a, b, d).

R. lateristriata has been initially described based on a series of specimens collected throughout the Western part of Java (Kottelat 2013) at type localities that are not included in the range distribution observed here (Fig. 5c). Most of the type localities belong to the watershed draining the largest urban areas in Java (i.e. Jakarta, Bogor, Bandung), which are highly impacted by anthropogenic activities. Previous observations highlighted that some localities in this part of Java were dominated by invasive species including *Xyphophorus* spp., *Oreochromis niloticus, Clarias gariepinus* and *Poecilia* spp. (Dahruddin et al. 2017), and very few of the sites visited in the western part of Java resulted in the capture of *Rasbora* specimens (Jawa Barat, Table S1). Also it is possible that we failed to capture *R. lateristriata* while the species was present, this result suggests that it has become rare, at least in this part of Java. The concordance between morphology-based and sequence-based species delimitation, however seems to confirm its validity.

Rasbora baliensis was reported as a species endemic to the crater lakes in Bali while *R. lateristriata* replaced it in the rivers of Bali (Kottelat et al. 1993; Kottelat 2013). Our study shows that *R. baliensis* has a range distribution surprisingly wider than expected as its presence is detected until East Java and Lombok (Fig. 5d). The confusion that reigns over the know range distribution of *R. baliensis* and *R. lateristriata* is surprising considering that they show non-overlapping numbers of scale at the lateral line, a character that has been previously overlooked. The otherwise morphological similarity between the two species is likely to account for the reported occurrence of *R. lateristriata* in most Lesser Sunda Islands including Bali, Lombok and Sumbawa (Kottelat et al. 1993). The disjunctive range distribution of both species, as well as the reciprocal monophyly that was captured by all DNA-based delimitation methods, argue that *R. baliensis* is a valid taxon. Furthermore, the presence of *R. baliensis* in east Java and Lombok is reported here for the first time while we show that *R. lateristriata* is restricted to central Java (Fig. 5d).

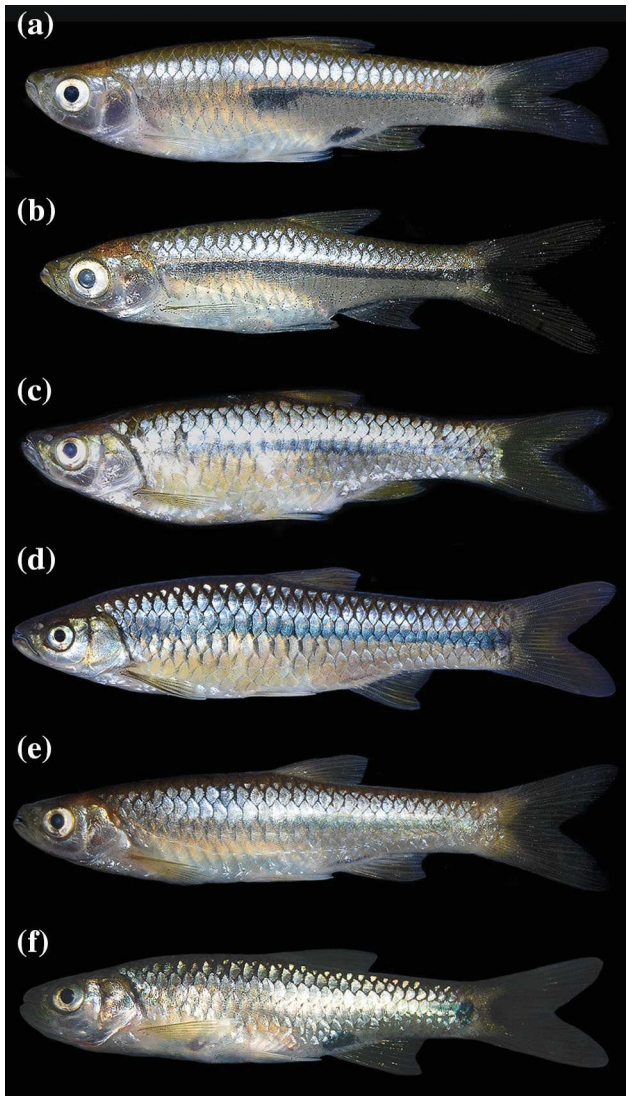

Fig. 4 Selected specimen photographs of each of the 6 *Rasbora* species collected and recognized in the present study. **a** *Rasbora aprotaenia* (specimen BIF1501; SL = 47 mm; Ci Siih, Banten, Java). **b** *Rasbora argyrotaenia* (specimen BIF976; SL = 33 mm; Cilacap, Central Java). **c** *Rasbora baliensis* (specimen BIF2351; SL = 72 mm; Jembrana, West Bali). **d** *Rasbora lateristriata* (specimen BIF3619; SL = 87 mm; Kali Dauwan, Mojokerto, East Java). **e** *Rasbora* sp1 (specimen BIF864; SL = 61 mm; Kali Pelus, Purwokerto, Central Java). **f** *Rasbora* sp2 (specimen BIF155; SL = 43 mm; Ci Heulang, Sukabumi, West Java)

Finally, two new taxa are discovered here, *Rasbora* sp1 and *Rasbora* sp2 (Figs. 2, 4). The COI tree suggests that both taxa are closely related to *R. aprotaenia*. The coloration pattern of *R.* sp1 is markedly different from *R. aprotaenia*, however as *R.* sp1 exhibit a thin dark line, instead of a dark spot in *R. aprotaenia*, at the proximal margin of the anal fin and a thin and a diffuse dark line along the lateral line (Fig. 4e). *Rasbora* sp2 is much more similar to *R. aprotaenia* in terms of coloration pattern, and the fact that a single

specimen was captured warrants further studies to enable its formal description. The diversification of this clade of closely related species is inferred to happen during the last 2.5 Ma in a restricted area of the western part of Java, a pattern previously described for several species largely distributed in Java such as *Barbodes binotatus, Channa gachua* and *Glyptothorax platypogon* (Hutama et al. 2017).

Taxonomy of *Nemacheilus* species in Java

The DNA-based delimitation schemes are consistent with the presence of two species in Java, namely *Nemacheilus fasciatus*, reported to occur in Java and Sumatra, and *Nemacheilus chrysolaimos*, an endemic species of Java (Kottelat et al. 1993; Kottelat 2013). The examination of the type locality of *N. fasciatus* provided a conflicting result with the initial identification performed based on the diagnostic morphological characters described in Kottelat (1993). The type locality of *N. chrysolaimos* is unknown and the type locality of *N. fasciatus* is located in the western part of Java (Fig. 6). Following Kottelat's monograph (1993), the Western species is *Nemacheilus chrysolaimos*. According to our study, however the *Nemacheilus* species of western Java is *N. fasciatus, N. chrysolaimos* being largely distributed in Central and Eastern Java (Fig. 6). A reexamination of the diagnostic morphological characters proposed to differentiate both *Nemacheilus* species in Java at the light of the present results indicates that *N. fasciatus* is the species that is distinguished in having a winged flap on the margin of the anterior naris (Fig. 7c, d), contrasting with an anterior naris valve pierced at the tip of a tube in *N. chrysolaimos* (Fig. 7a, b). The coloration patterns of each species are also distinct with *N. chrysolaimos* presenting 14–18 dark blotches along lateral line with 11–12 dark saddles across the back contrasting with *N. fasciatus* exhibiting 9–18 dark bars of irregular shape (Kottelat et al. 1993). The identity of both species was further confirmed by the examination of the type specimens of both species (*N. chrysolaimos*, B2972, B3961; *N fasciatus*, B2798) deposited at the MNHN Paris, as well as the original illustration of *N. chrysolaimos* depicting a coloration pattern consisting of irregular saddles, consistent with the coloration pattern of the specimens assigned to *N. chrysolaimos* in the present study.

Species distribution and conservation genetics

All the species under study here were previously reported as endemic species of the islands of Java, Bali and Lombok.

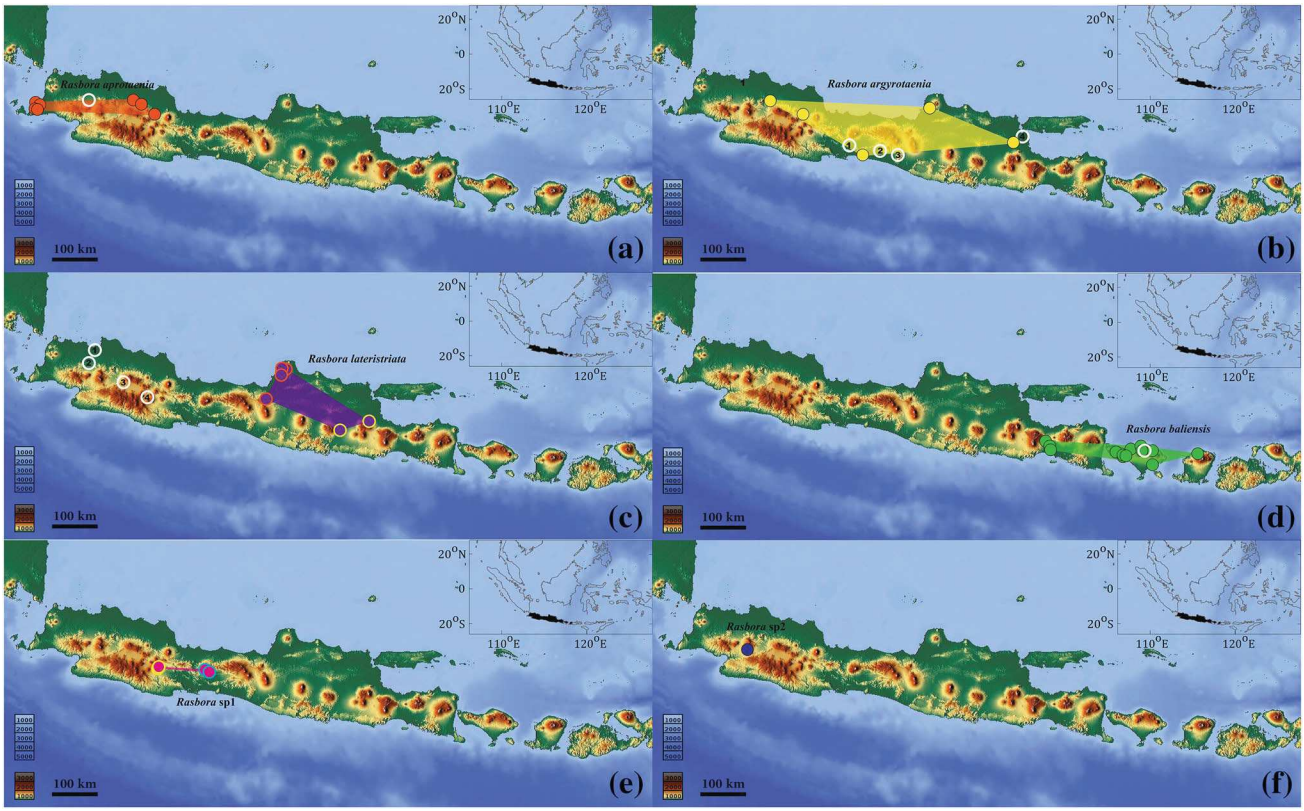

Fig. 5 Species range distribution of the 6 *Rasbora* species recognized in the present study. Colored dots represent collection sites. White circles represent type localities. **a** *Rasbora aprotaenia*. **b** *Rasbora argyrotaenia*. **c** *Rasbora lateristriata*. **d** *Rasbora baliensis*. **e** *Rasbora* sp1. **f** *Rasbora* sp2

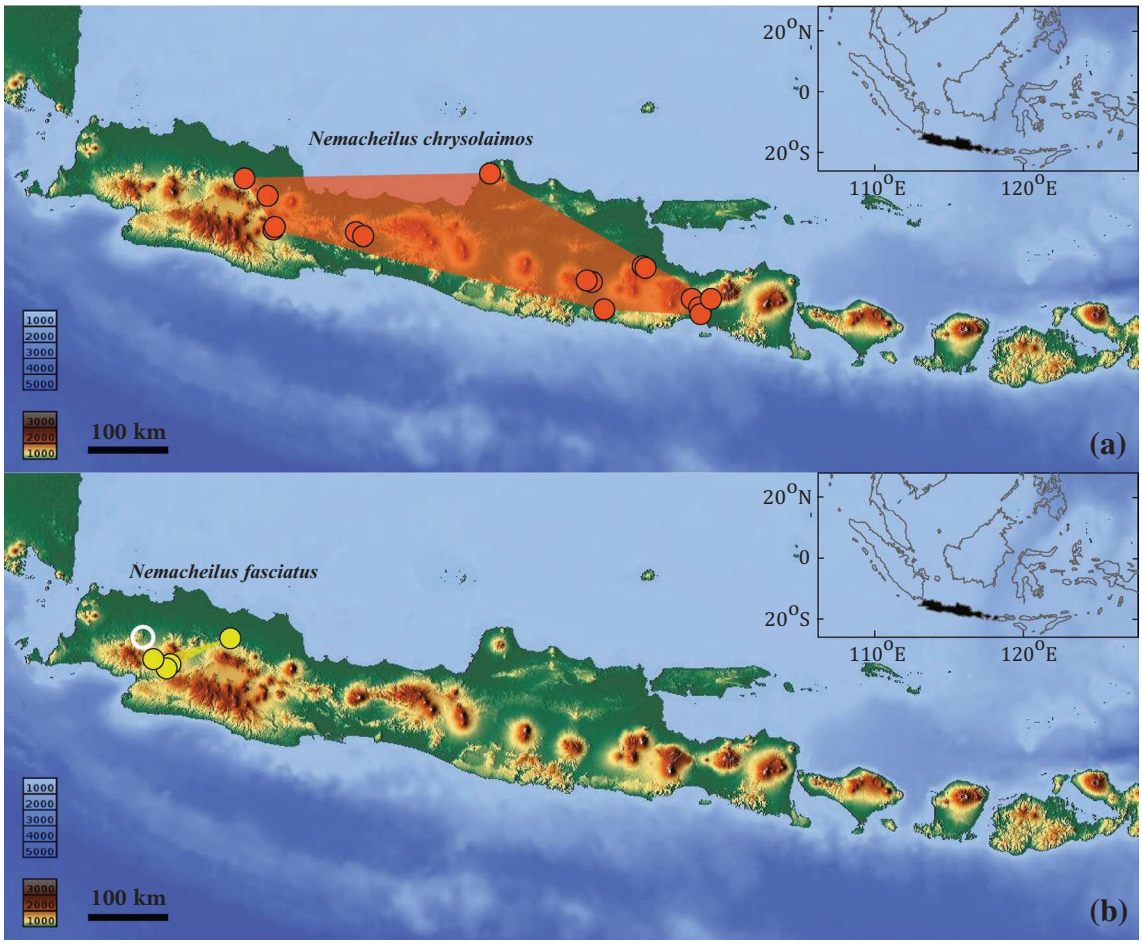

Fig. 6 Species range distribution of the 2 *Nemacheilus* species recognized in the present study. Colored dots represent collection sites. White circles represent type localities. **a** *Nemacheilus chrysolaimos*. **b** *Nemacheilus fasciatus*

The present study highlights that most of the species examined here have range distribution much more restricted than previously acknowledged. Excepting *N. chrysolaimos* and *R. argyrotaenia* detected in east and central Java, most species are confined to a restricted set of watersheds. A clear separation is observed between the eastern versus the central and east Java lineages, a pattern previously reported among cryptic lineages within widespread species in Java such as *Barbodes binotatus, Glyptothorax platypogon* and *Channa gachua* (Dahruddin et al. 2017; Hutama et al. 2017). This biogeographic transition between East and West watersheds in Java is related to the ontogeny of Java, resulting from the merging of two volcanic arches, one is the west and a second in the east (including East Java and Bali), that emerged between 10 and 5 Ma and further aggregated during the last 5 million years (Lohman et al. 2011). This scenario is reflected in *Rasbora* with the range distribution of *R. baliensis* encompassing the eastern part of Java, Bali and Lombok. A similar trans-distribution across the Bali strait was previously observed in *Barbodes binotatus* and *Channa gachua* (Hutama et al. 2017). Divergence levels between the western most species and the central/east species are surprisingly higher than previously observed in Java with a divergence between *N. fasciatus* and *N. chrysolaimus* inferred at around 10 Ma (1.01–17.90 Ma, 95% HPD) and the divergence of *R. argyrotaenia* from other *Rasbora* inferred at around 11.5 Ma (10.61–12.55 Ma, 95% HPD). These age estimates are much older than reported for other widespread species in Java with species MRCA estimated to occur around 3 Ma for several species (Hutama et al. 2017). The age of the MRCA of sampled species of *Nemacheilus* and *Rasbora* also trace back beyond the ontogeny of Java, suggesting that the evolutionary history of these Javanese species started prior to the rise of this island, during the initial settlement of Sundaland (Lohman et al. 2011). These old splits particularly contrast in *Rasbora* with divergence age estimates for other *Rasbora* species that do not exceed 3.5 Ma, an age consistent with previous age estimates of the MRCA of the coalescent trees of several widespread species in Java (Hutama et al. 2017). Interestingly, *N. fasciatus* and *R. argyrotaenia* are both cited

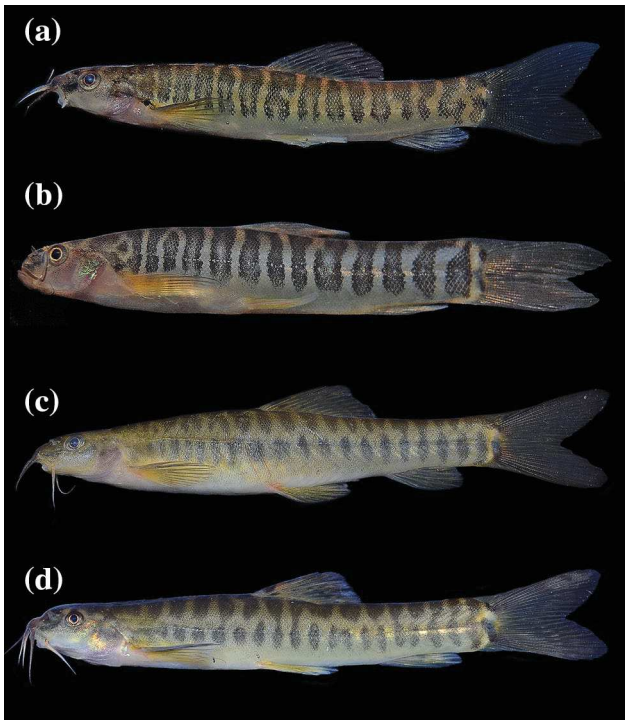

Fig. 7 Selected specimen photographs of each of the 2 *Nemacheilus* species collected and recognized in the present study. **a** *Nemacheilus fasciatus* (specimen BIF495; SL = 42 mm; Ci Asem, Purwakarta, West Java). **b** *Nemacheilus fasciatus* (specimen BIF163; SL = 45 mm; Ci Heulang, Central Java). **c** *Nemacheilus chrysolaimos* (specimen BIF2032; SL = 58 mm; Ngerjo, Blitar, East Java). **d** *Nemacheilus chrysolaimos* (specimen BIF2074; SL = 55 mm; Bicoro, Lumajang, East Java)

from Sumatra suggesting a recent colonization of Java from Sumatra for those species. This scenario was recently suggested for *Rasbora* based on a phylogenetic assessment of Java species based on mitochondrial genomes (Kusuma et al. 2016), and highlights the different origins of the western and eastern species of *Nemacheilus* and *Rasbora*. The *Rasbora* tree topology, supporting close relationships between central and east Java species, and age estimates, below 3.5 Ma, suggest that most *Rasbora* species diversified through in-situ speciation in Java, a hypothesis previously suggested (Kusuma et al. 2016).

The general trend of highly restricted range distribution, related to a hypothesized origin though in-situ diversification in Java and Bali for *Rasbora*, highlights the sensitive conservation status of most endemic species in Java. The distribution patterns observed here are highly fragmented with little overlap between range distributions of sister-species, a pattern that was expected considering the volcanic origin of Java, Bali and Lombok islands during the Quaternary (Lohman et al. 2011). Genetic diversity patterns are consistent with these observations as nucleotidic diversity is generally low for *Rasbora* and *Nemacheilus* species, a molecular signature also consistent with a recent origin through in-situ diversification. Cryptic diversity is also reported here for *R. lateristriata* and *R.* sp1, with cryptic lineages diverging around 1 Ma for both species, further accentuating the fragmented status of the populations of these species. As such, most *Rasbora* species should be considered as complexes of small populations, highly sensitive to further reductions in population size that may eventually lead to a vortex of extinction (Gilpin and Soulé 1986; Fagan and Holmes 2006).

Java is the most densely populated island in Sundaland and anthropogenic activities have an alarming impact on the rivers of this region (Spracklen et al. 2015; Breckwoldt et al. 2016; Hayati et al. 2017; Garg et al. 2018). Severe reductions of aquatic habitat size during the last two decades have been reported with significant impacts of water pollution generated by industrial wastes and agricultural run-offs, but also the introduction of exotic species (Eidman 1989; Dahruddin et al. 2017). Considering the very narrow range distributions observed here for most endemic species and the concerning environmental context of Java, this study stresses that conservation efforts even at a small spatial scale could have significant impacts in the conservation of these endemic species. Translocations of livestock are common in Indonesia and are often used as a management measure (Dahruddin et al. 2017). The presence of cryptic diversity, newly discovered, yet undescribed, taxa and their close morphological affinities make accurate species identifications difficult. In this context, the introduction of multiple species through translocation programs due to species misidentifications are likely if not guided by DNA barcoding. Considering the young age of most of the species examined here and their allopatric distributions, the effectiveness of reproductive isolation mechanisms after secondary contact is questionable and the occurrence of hybridization cannot by discarded. Both genera have become scarce in the Western part of Java. Combined with the low genetic diversity at the nucleotide level, restoration program through genetic rescue (i.e. increasing population fitness through the repletion of genetic diversity by immigration) are probably required (Whiteley et al. 2014). If such translocation programs were to be implemented in the future, we advocate for short spatial scale translocations in order to avoid secondary contact among closely related species and identify immigrants, to the species level using DNA barcodes.

Conclusion

As in previous molecular studies in the area, the present study highlights gaps of knowledge in the taxonomy and distribution of the freshwater fishes of Java and Bali, and further highlights the complexity of diversity patterns in this part of Sundaland. This assessment of *Rasbora* and

Nemacheilus diversity through DNA barcodes sheds light on the species biological status and distribution in South Sundaland and warrants further studies. In particular, the two new taxa discovered here need additional sampling efforts to be accurately described. The present study highlights the sensitive status of most species owing to their restricted range and low genetic diversity.

Acknowledgements The authors wish to thank Siti Nuramaliati Prijono, Bambang Sunarko, Witjaksono, Mohammad Irham, Marlina Adriyani, Ruliyana Susanti, Rosichon Ubaidillah, Hari Sutrisno and Muhamad Syamsul Arifin Zein at Research Centre for Biology (RCB-LIPI); Jean-Paul Toutain, Robert Arfi, Valérie Verdier and Jean-François Agnèse from the 'Institut de Recherche pour le Développement'; Joel Le Bail and Nicolas Gascoin at the French embassy in Jakarta for their continuous support. We are thankful Sumanta at IRD Jakarta for his help during the field sampling. Part of the present study was funded by the Institut de Recherche pour le Développement (UMR226 ISE-M and IRD through incentive funds), the MNHN (UMR BOREA), the RCB-LIPI, the French Ichthyological Society (SFI), the Foundation de France and the French embassy in Jakarta. The Indonesian Ministry of Research and Technology approved this study and field sampling was conducted according to the research permits 097/SIP/FRP/SM/IV/2014 for Philippe Keith, 60/EXT/SIP/FRP/SM/XI/2014 for Frédéric Busson and 41/EXT/SIP/FRP/SM/VIII/2014 for Nicolas Hubert. Sequence analysis was aided by funding from the government of Canada through Genome Canada and the Ontario Genomics Institute in support of the International Barcode of Life project. We thank Paul Hebert, Robert Hanner and Evgeny Zakharov as well as BOLD and CCDB staffs at the University of Guelph for their valuable support. Finally, we thank Anti Vasemägi and the three anonymous reviewers for providing constructive comments of earlier versions of the manuscript. This publication has ISEM Number 2018-279-SUD.

References

April J, Mayden R, Hanner L, Bernatchez RH L (2011) Genetic calibration of species diversity among North America's freshwater fishes. Proc Natl Acad Sci USA 108:10602–10607

Arhens D, Fujisawa T, Krammer HJ, Eberle J, Fabrizi S, Vogler AP (2016) Rarity and incomplete sampling in DNA-based species delimitation. Syst Biol 65:478–494

Avise JC (1989) Molecular markers, natural history and evolution. Chapman and Hall, New York

Bermingham E, McCafferty S, Martin AP (1997) Fish biogeography and molecular clocks: perspectives from the Panamanian isthmus. In: Kocher TD, Stepien CA (eds) Molecular systematics of fishes. CA Academic Press, San Diego, pp 113–128

Blair C, Bryson JRW (2017) Cryptic diversity and discordance in single-locus species delimitation methods within horned lizards (Phrynosomatidae: Phrynosoma). Mol Ecol Resour 17:1168–1182

Bouckaert RR, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10:e1003537

Breckwoldt A, Dsikowitzky L, Baum G, Ferse SCA, van der Wulp S, Kusumanti I, Ramadhan A, Adrianto L (2016) A review of stressors, uses and management perspectives for the larger Jakarta Bay Area, Indonesia. Mar Pollut Bull 110:790–794

Brown SDJ, Collins RA, Boyer S, Lefort C, Malumbres-Olarte J, Vink CJ, Cruickshank RH (2012) SPIDER: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Mol Ecol Resour 12:562–565

Dahruddin H, Hutama A, Busson F, Sauri S, Hanner R, Keith P, Hadiaty RK, Hubert N (2017) Revisiting the ichthyodiversity of Java and Bali through DNA barcodes: taxonomic coverage, identification accuracy, cryptic diversity and identification of exotic species. Mol Ecol Resour 17:288–299

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772

Durand JD, Hubert N, Shen KN, Borsa P (2017) DNA barcoding grey mullets. Rev Fish Biol Fisheries 27:233–243

Eidman HM (1989) Exotic aquatic species introduction into Indonesia, vol 3. Exotic aquatic organisms in Asia Asian Fisheries Society Special Publication, Bethesda, pp 57–62

Eschmeyer WN, Fricke R, van der Laan R (2018) Catalog of fishes electronic version. Accessed March 2018

Ezard T, Fujisawa T, Barraclough TG (2009) Splits: SPecies' LImits by Threshold Statistics. R package version 1.0-18/r45. Available from http://R-Forge.R-project.org/projects/splits/

Fagan WF, Holmes EE (2006) Quantifying the extinction vortex. Ecol Lett 9:51–60

Froese R, Pauly D (2014) FishBase. World Wide Web electronic publication. http://www.fishbase.org

Fujisawa T, Barraclough TG (2013) Delimiting species using single-locus data and the generalized mixed Yule coalescent approach: a revised method and evaluation on simulated data sets. Syst Biol, 62:707–24

Garg T, Hamilton SE, Hochard JP, Kresch EP, Talbot J (2018) (Not so) gently down the stream: River pollution and health in Indonesia. J Environ Econ Manag 92:35–53

Garnett ST, Christidis L (2017) Taxonomy anarchy hampers conservation. Nature 546:25–27

Geiger MF, Herder F, Monaghan MT, Almada V, Barbieri R, Bariche M, Berrebi, Bohlen P, Casal-Lopez M, Delmastro GB (2014) Spatial heterogeneity in the mediterranean biodiversity hotspot affects barcoding accuracy of its freshwater fishes. Mol Ecol Resour 14:1210–1221

Gilpin E, Soulé M (1986) Minimum viable populations: processes of species extinction. In: Soulé ME (ed) Conservation biology: the science of scarcity and diversity. Sinauer, Sunderland. pp 19–34

Guindon S, Gascuel O (2003) A simple, fast and Accurate algorithm to estimate large phylogenies by Maximum Likelihood. Syst Biol 52:696–704

Hayati A, Tiantono N, Mirza MF, Putra IDS, Abdizen MM, Seta AR, Solikha BM, Fu'adil MH, Putranto TWC, Affandi M, Rosmaninda (2017) Water quality and fish diversity in the Brantas river, East Java, Indonesia. J Biol Res 22:43–49

Hoffman M, Hilton-Taylor C, Angulo A, Böhm M, Brooks TM, Butchart SHM, Carpenter KE, Chanson J, Collen B, Cox NA et al (2010) The impact of conservation on the status of the world's vertebrates. Science 330:1503–1509

Hubert N, Hanner R (2015) DNA barcoding, species delineation and taxonomy: a historical perspective. DNA Barcodes 3:44–58

Hubert N, Hanner RH, Holm E, Mandrak NE, Taylor EB, Burridge M, Watkinson DA, Dumont P, Curry A, Bentzen P, Zhang J, April J, Bernatchez L (2008) Identifying Canadian freshwater fishes through DNA barcodes. PLoS ONE, 3:e2490

Hubert N, Meyer C, Bruggemann JH, Guérin F, Komeno RJL, Espiau B, Causse R, Williams JT, Planes S (2012) Cryptic diversity in Indo-Pacific coral reef fishes revealed by DNA-barcoding provides new support to the centre-of-overlap hypothesis. PLoS ONE, 7:e28987

Hubert N, Wibowo A, Busson F, Caruso D, Sulandari S, Nafiqoh N, Rüber L, Pouyaud L, Avarre JC, Herder F, Hanner R, Keith P, Hadiaty RK (2015) DNA barcoding Indonesian freshwater fishes: challenges and prospects. DNA Barcodes 3:144–169

Hubert N, Dettai A, Pruvost P, Cruaud C, Kulbicki M, Myers RF, Borsa P (2018) Geography and life history traits account for the accumulation of cryptic diversity among Indo-West Pacific coral reef fishes. Mar Ecol Prog Ser 583:179–193

Hutama A, Dahruddin H, Busson F, Sauri S, Keith P, Hadiaty RK, Hanner R, Suryobroto B, Hubert N (2017) Identifying spatially concordant evolutionary significant units across multiple species through DNA barcodes: application to the conservation genetics of the freshwater fishes of Java and Bali. Global Ecol Conserv 12:170–187

Ivanova NV, Zemlak TS, Hanner RH, Hébert PDN (2007) Universal primers cocktails for fish DNA barcoding. Mol Ecol Notes 7:544–548

Jaafar TNAM, Taylor MI, Mohd Nor SA, De Bruyn M, Carvalho GR (2012) DNA barcoding reveals cryptic diversity within commercially exploited Indo-Malay carangidae (Teleosteii: Perciformes). PLoS ONE, 7:e49623

Kadarusman HN, Hadiaty RK, Paradis E, Pouyaud L (2012) Cryptic diversity in Indo-Australian rainbowfishes revealed by DNA Barcoding: implications for conservation in a biodiversity hotspot candidate. Plos ONE 7:e40627

Kapli P, Zhang SL, Kobert J, Pavlidis K, Stamatakis P, Flouri A T (2017) Multi-rate Poisson Tree Processes for single-locus species delimitation under Maximum Likelihood and Markov Chain Monte Carlo. Bioinformatics 33:1630–1638

Kekkonen M, Hebert PDN (2014) DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Mol Ecol Resour 14:706–715

Kekkonen M, Mutanen M, Kaila L, Nieminen M, Hebert PDN (2015) Delineating species with DNA barcodes: a case of taxon dependent method performance in moths. PLoS ONE 10:e0122481

Knebelsberger T, Dunz AR, Neumann D, Geiger MF (2015) Molecular diversity of Germany's freshwater fishes and lampreys assessed by DNA barcoding. Mol Ecol Resour 15:562–572

Kottelat M (2013) The fishes of the inland waters of Southeast Asia: a catalog and core bibliography of the fishes known to occur in freshwaters, mangroves and estuaries. Raffles Bull Zool Suppl 27:1–663

Kottelat M, Whitten AJ, Kartikasari SR, Wirjoatmodjo S (1993) Freshwater fishes of western indonesia and sulawesi. Periplus editions, Singapore

Kusuma WE, Ratmuangkhwang S, Kumazawa Y (2016) Molecular phylogeny and historical biogeography of the Indonesian freshwater fish Rasbora lateristriata species complex (Actinopterygii: Cyprinidae): cryptic species and west-to-east divergences. Mol Phylogenet Evol 105:212–223

Lamoreux JF, Morrison JC, Ricketts TH, Olson DM, Dinerstein E, McKnight M, Shugart HH (2006) Global tests of biodiversity concordance and the importance of endemism. Nature 440:212–214

Lohman K, De Bruyn M, Page T, Von Rintelen K, Hall R, Ng PKL, Shih H-T, Carvalho GR, Von Rintelen T (2011) Biogeography of the Indo-Australian archipelago. Annu Rev Ecol Evol Syst 42:205–226

Machado VN, Collins RA, Ota RP, Andrade MC, Farias IP, Hrbek T (2018) One thousand DNA barcodes of piranhas and pacus reveal geographic structure and unrecognised diversity in the Amazon. Sci Rep 8:8387

Miralles A, Vences M (2013) New metrics for comparison of taxonomies eveal striking discrepancies among species delimitation methods in Madascincus lizards. PLoS ONE 8:e68242

Moritz C (1994) Defining 'Evolutionary Significant Units' for conservation. Trends Ecol Evol 9:373–375

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. Nature 403:853–858

Paradis E (2010) pegas: an R package for poluation genetics with an integrated modular approach. Bioinformatics 26:419–420

Paradis E, Claude J, Strimmer K (2004) ape: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290

Pereira LHG, Hanner R, Foresti F, Oliveira C (2013) Can DNA barcoding accurately discriminate megadiverse Neotropical freshwater fish fauna ? BMC Genet 14:20

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst Biol 55:595–606

Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, automatic barcode gap discovery for primary species delimitation. Mol Ecol 21:1864–1877

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ratnasingham S, Hebert PDN (2007) BOLD: the barcode of life data system. Mol Ecol Notes 7:355–364

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the barcode index number (BIN) system. PLoS ONE 8:e66213

Spracklen DV, Reddington CL, Gaveau DLA (2015) Industrial concessions, fires and air pollution in Equatorial Asia. Environ Res Lett 10:091001

Steinke D, Hanner R (2011) The FISH-BOL collaborators' protocol. Mitochondrial DNA 22:10–14

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729

Vogler AP, DeSalle R (1994) Diagnosing units of conservation management. Conserv Biol 6:170–178

Ward RD, Hanner RH, Hebert PDN (2009) The campaign to DNA barcode all fishes, FISH-BOL. J Fish Biol 74:329–356

Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA (2014) Genetic rescue to the rescue. Trends Ecol Evol 30:42–49

Winemiller KO, McIntyre PB, Castello L, Fluet-Chouinard E, Giarrizzo T, Nam S, Baird IG, Darwall W, Lujan NK, Harrison I, Stiassny MLJ, Silvano RAM, Fitzgerald DB, Pelicice FM, Agostinho AA, Gomes LC, Albert JS, Baran E, Petrere M, Zarfl C, Mulligan M, Sullivan JP, Arantes CC, Sousa LM, Koning AA, Hoeinghaus DJ, Sabaj M, Lundberg JG, Armbruster J, Thieme ML, Petry P, Zuanon J, Vilara GT, Snoeks J, Ou C, Rainboth W, Pavanelli CS, Akama A, Soesbergen AV, Sáenz L (2016) Balancing hydropower and biodiversity in the Amazon, Congo, and Mekong. Science 351:128–129

Winterbottom R, Hanner R, Burridge M, Zur M (2014) A cornucopia of cryptic species—a DNA barcode analysis of the gobiid genus Trimma (Percomorpha, Gobiiformes). Zookeys 381:79–111

Woodruff DS (2010) Biogeography and conservation in Southeast Asia: how 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugium-phase biodiversity. Biodivers Conserv 19:919–941

Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876

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

Appendix B
Disentangling the Taxonomy of the Subfamily Rasborinae (Cypriniformes, Danionidae) in Sundaland Using DNA Barcodes

*Arni Sholihah[1,2], Erwan Delrieu-Trottin[2,3], Tedjo Sukmono[4], Hadi Dahruddin[2,5], Renny Risdawati[6], Roza Elvyra[7], Arif Wibowo[8,9], Kustiati Kustiati[10], Frédéric Busson[2,11], Sopian Sauri[5], Ujang Nurhaman[5], Edmond Dounia[12], Muhamad Syamsul Arifin Zein[5], Yuli Fitriana[5], Ilham Vemendra Utama[5], Zainal Abidin Muchlisin[13], Jean-François Agnèse[2], Robert Hanner[14], Daisy Wowor[5], Dirk Steinke[14], Philippe Keith[11], Lukas Rüber[15,16] & Nicolas Hubert[2]**

1 Institut Teknologi Bandung, School of Life Sciences and Technology, Bandung, Indonesia.
2 UMR 5554 ISEM (IRD, UM, CNRS, EPHE), Université de Montpellier, Place Eugène Bataillon, 34095, Montpellier cedex 05, France.
3 Museum für Naturkunde, Leibniz-Institut für Evolutions und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Invalidenstrasse 43, Berlin, 10115, Germany.
4 Universitas Jambi, Department of Biology, Jalan Lintas Jambi - Muara Bulian KM 15, 36122, Jambi, Sumatra, Indonesia.
5 Division of Zoology, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Jalan Raya Jakarta Bogor KM 46, Cibinong, 16911, Indonesia.
6 Department of Biology Education, STKIP PGRI Sumatera Barat, Jl Gunung Pangilun, Padang, 25137, Indonesia.
7 Universitas Riau, Department of Biology, Simpang Baru, Tampan, Pekanbaru, 28293, Indonesia.
8 Southeast Asian Fisheries Development Center, Inland Fisheries Resources Development and Management Department, 8 Ulu, Seberang Ulu I, Palembang, 30267, Indonesia.
9 Research Institute for Inland Fisheries and Fisheries extensions, Agency for Marine and Fisheries Research, Ministry of Marine Affairs and Fisheries., Jl. H.A. Bastari No. 08, Jakabaring, Palembang, 30267, Indonesia.
10 Universitas Tanjungpura, Department of Biology, Jalan Prof. Dr. H. Hadari Nawawi, Pontianak, 78124, Indonesia.
11 UMR 7208 BOREA (MNHN-CNRS-UPMC-IRD-UCBN), Muséum National d'Histoire Naturelle, 43 rue Cuvier, 75231, Paris cedex 05, France.
12 UMR 5175 CEFE (IRD, UM, CNRS, EPHE), 1919 route de Mende, 34293, Montpellier cedex 05, France.
13 Syiah Kuala University, Faculty of Marine and Fisheries, Banda, Aceh, 23111, Indonesia.
14 Department of Integrative Biology, Centre for Biodiversity Genomics, 50 Stone Rd E, Guelph, ON, N1G2W1, Canada.
15 Naturhistorisches Museum Bern, Bernastrasse 15, Bern, 3005, Switzerland.
16 Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, 3012, Bern, Switzerland.
*email: nicolas.hubert@ird.fr

Scientific Reports volume 10, article number: 2818 (2020)
https://doi.org/10.1038/s41598-020-59544-9
Received: 21 November 2019; Accepted: 20 January 2020;
Published online: 18 February 2020

SCIENTIFIC REPORTS
nature research

OPEN

Disentangling the taxonomy of the subfamily Rasborinae (Cypriniformes, Danionidae) in Sundaland using DNA barcodes

Arni Sholihah[1,2], Erwan Delrieu-Trottin[2,3], Tedjo Sukmono[4], Hadi Dahruddin[2,5], Renny Risdawati[6], Roza Elvyra[7], Arif Wibowo[8,9], Kustiati Kustiati[10], Frédéric Busson[2,11], Sopian Sauri[5], Ujang Nurhaman[5], Edmond Dounias[12], Muhamad Syamsul Arifin Zein[5], Yuli Fitriana[5], Ilham Vemendra Utama[5], Zainal Abidin Muchlisin[13], Jean-François Agnèse[2], Robert Hanner[14], Daisy Wowor[5], Dirk Steinke[14], Philippe Keith[11], Lukas Rüber[15,16] & Nicolas Hubert[2*]

Sundaland constitutes one of the largest and most threatened biodiversity hotspots; however, our understanding of its biodiversity is afflicted by knowledge gaps in taxonomy and distribution patterns. The subfamily Rasborinae is the most diversified group of freshwater fishes in Sundaland. Uncertainties in their taxonomy and systematics have constrained its use as a model in evolutionary studies. Here, we established a DNA barcode reference library of the Rasborinae in Sundaland to examine species boundaries and range distributions through DNA-based species delimitation methods. A checklist of the Rasborinae of Sundaland was compiled based on online catalogs and used to estimate the taxonomic coverage of the present study. We generated a total of 991 DNA barcodes from 189 sampling sites in Sundaland. Together with 106 previously published sequences, we subsequently assembled a reference library of 1097 sequences that covers 65 taxa, including 61 of the 79 known Rasborinae species of Sundaland. Our library indicates that Rasborinae species are defined by distinct molecular lineages that are captured by species delimitation methods. A large overlap between intraspecific and interspecific genetic distance is observed that can be explained by the large amounts of cryptic diversity as evidenced by the 166 Operational Taxonomic Units detected. Implications for the evolutionary dynamics of species diversification are discussed.

[1]Instut Teknologi Bandung, School of Life Sciences and Technology, Bandung, Indonesia. [2]UMR 5554 ISEM (IRD, UM, CNRS, EPHE), Université de Montpellier, Place Eugène Bataillon, 34095, Montpellier, cedex, 05, France. [3]Museum für Naturkunde, Leibniz-Institut für Evolutions und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Invalidenstrasse 43, Berlin, 10115, Germany. [4]Universitas Jambi, Department of Biology, Jalan Lintas Jambi - Muara Bulian Km15, 36122, Jambi, Sumatra, Indonesia. [5]Division of Zoology, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Jalan Raya Jakarta Bogor Km 46, Cibinong, 16911, Indonesia. [6]Department of Biology Education, STKIP PGRI Sumatera Barat, Jl Gunung Pangilun, Padang, 25137, Indonesia. [7]Universitas Riau, Department of Biology, Simpang Baru, Tampan, Pekanbaru, 28293, Indonesia. [8]Southeast Asian Fisheries Development Center, Inland Fisheries Resources Development and Management Department, 8 Ulu, Seberang Ulu I, Palembang, 30267, Indonesia. [9]Research Institute for Inland Fisheries and Fisheries extensions, Agency for Marine and Fisheries Research, Ministry of Marine Affairs and Fisheries., Jl. H.A. Bastari No. 08, Jakabaring, Palembang, 30267, Indonesia. [10]Universitas Tanjungpura, Department of Biology, Jalan Prof. Dr. H. Hadari Nawawi, Pontianak, 78124, Indonesia. [11]UMR 7208 BOREA (MNHN-CNRS-UPMC-IRD-UCBN), Muséum National d'Histoire Naturelle, 43 rue Cuvier, 75231, Paris, cedex, 05, France. [12]UMR 5175 CEFE (IRD, UM, CNRS, EPHE), 1919 route de Mende, 34293, Montpellier, cedex, 05, France. [13]Syiah Kuala University, Faculty of Marine and Fisheries, Banda, Aceh, 23111, Indonesia. [14]Department of Integrative Biology, Centre for Biodiversity Genomics, 50 Stone Rd E, Guelph, ON, N1G2W1, Canada. [15]Naturhistorisches Museum Bern, Bernastrasse 15, Bern, 3005, Switzerland. [16]Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, 3012, Bern, Switzerland. *email: nicolas.hubert@ird.fr

Over the past two decades, the spectacular aggregation of species in biodiversity hotspots has attracted attention by scientists and stakeholders alike[1–4]. However, this exceptional concentration of often-endemic species at small spatial scales is threatened by the rise of anthropogenic disturbances. Of the 26 initially identified terrestrial biodiversity hotspots[1], the ones located in Southeast Asia (SEA) (Indo-Burma, Philippines, Sundaland and Wallacea) currently rank among the most important both in terms of species richness and the extend of endemism but also rank as the most threatened by human activities[3]. Sundaland is currently the most diverse terrestrial biodiversity hotspot of SEA and is the most threatened[5]. Sundaland comprises Peninsular Malaysia and the islands of Java, Sumatra, Borneo, and Bali and its diversity originates from the complex geological history of the region, linked to major tectonic changes in the distribution of land and sea during the last 50 Million years (My)[6], but also from eustatic fluctuations that have sporadically connected and disconnected Sundaland landmasses during glacial-interglacial cycles in the Pleistocene[7–9]. Therefore, Sundaland biogeography has received increased attention over the past decade resulting in the detection of contrasting spatial and temporal patterns in various groups[9–14].

Species richness within vertebrate groups is high in Sundaland[1] and freshwater fishes are no exceptions to that. With more than 900 species reported to date, and with nearly 45 percent of endemism, Sundaland's ichthyofauna is the largest in SEA and accounts for nearly 75 percent of the entire ichthyodiversity of the Indonesian archipelago[15]. The inventory of Sundaland's freshwater fishes started more than two centuries ago and despite the acceleration of species discovery over the last three decades, it is still a work in progress[15]. The complexity of this inventory was partly exacerbated by the abundance of minute species *i.e.* less than 5 cm length[15], but also by the inconsistent use of species names through time for old descriptions due to the loss of type specimens or uncertainties in the location of type localities[16,17]. The family Cyprinidae *sensu lato* is a particularly good example for the complexity of Sundaland freshwater fishes taxonomy and systematics. The systematics of this large family of Cypriniformes, with over 3,000 species, has been controversial for more than a century[18]. Based on recent molecular phylogenetic studies[19–21], Tan and Armbruster[22] proposed a new classification dividing the Cyprinidae *sensu lato* into 12 families. The subfamily Rasborinae (Cypriniformes, Cyprinoidei, Danionidae) comprises roughly 140 species in 11 genera: *Amblypharyngodon*, *Boraras*, *Brevibora*, *Horadandia*, *Kottelatia*, *Pectenocypris*, *Rasbora*, *Rasboroides*, *Rasbosoma*, *Trigonopoma*, and *Trigonostigma*[22]. In Sundaland the subfamily Rasborinae is represented by 79 species in 7 genera. The genera *Amblypharyngodon*, *Horadandia*, *Rasboroides*, and *Rasbosoma* do not occur in Sundaland. By far the most species rich rasborine genus is *Rasbora* with over 100 species in total and 65 species in Sundaland. Long considered a catch-all group, several attempts have been made to provide a classification of the genus *Rasbora* that reflects phylogeny. In a comprehensive revision, Brittan[23] recognized 3 subgenera (*Rasbora*, *Rasboroides*, and *Megarasbora*) and divided *Rasbora* into 8 species complexes, now regarded as species groups[24] (Fig. 1). Subsequent authors have erected several new genera or suggested new species composition for the various *Rasbora* species groups[19,24–26]. Clearly, to better understand the evolutionary history of this unique group, the taxonomy and systematic of the Rasborinae needs to be better understood.

The use of standardized DNA-based approaches to the inventory of Sundaland's ichthyofauna resulted in the detection of considerable knowledge gaps[16,17,27]. In addition, substantial levels of cryptic diversity (*i.e.* morphologically unrecognized diversity) were repeatedly reported for a wide range of Sundaland freshwater fish taxa[10,27–33] including the Rasborinae[16]. The taxonomy of most Rasborinae species, particularly so for the genus *Rasbora*, remains challenging due to the diversity of the group and the morphological similarity of many closely related species. As a consequence, the actual distribution ranges of many species of Rasborinae are not well known.

This study aims to re-examine Rasborinae diversity in Sundaland. We generated a DNA barcode reference library to (1) explore biological species boundaries with DNA-based species delimitation methods, (2) validate species identity, taxonomy and precise range distribution by producing DNA barcodes from type localities or neighboring watersheds, (3) validate or revise of the previously published DNA barcodes records for the subfamily Rasborinae available on GenBank.

Results

Sequencing of the DNA barcode marker Cytochrome Oxidase 1 (COI) yielded a total of 991 new sequences (Table S2) from 189 sampling sites distributed across Sundaland (Fig. 2). Together with 106 DNA barcodes mined from GenBank and BOLD (Table S3), we assembled a DNA barcode reference library of 1,097 sequences from 65 taxa of Rasborinae and 1 taxon of Sundadanionidae (*Sundadanio retiarius*). The number of specimens analyzed per species ranged from 1 to 143, with an average of 14.6 sequences per species and only six species represented by a single sequence. The sequences ranged from 459 bp to 651 bp long, with 99 percent of the sequences being above 500 bp length, and no stop codons were detected, suggesting that all the sequences correspond to functional mitochondrial COI sequences. DNA barcodes for 61 of the 79 nominal species of Rasborinae reported from Sundaland were recovered (approximately 78%) corresponding to the 7 Rasborinae genera currently recognized (Table S1). The present study achieved a complete coverage at the species level for the genera *Boraras* (2 species), *Brevibora* (3 species), *Kottelatia* (1 species), *Trigonopoma* (2 species) and *Trigonostigma* (3 species). In turn, two out of the three *Pectenocypris* species (66%) and 48 out of the 65 *Rasbora* species (74%) currently recognized in Sundaland were collected (Table S1). Geographically, our dataset includes 86% of the Rasborinae of Borneo (38 out of 44), all the *Rasbora* species of Java (4 species) and 68% of the Rasborinae species of Sumatra (26 out of 38) were collected (Table S1). Finally, four undescribed taxa are highlighted, two taxa of *Rasbora* in Java, one taxon of *Trigonostigma* in Borneo (Table S2) and an additional *Rasbora* taxon, previously assigned to *R. paucisqualis* in the literature (Table S3).

Species delimitation analyses provided varying numbers of Operational Taxonomic Units (OTUs) among methods (Fig. 3): 129 for PTP, 95 for mPTP, 178 for GMYC, 191 for mGMYC, 175 for ABGD and 146 for RESL (Table S3). Our consensus delimitation scheme yielded 166 OTUs, including 165 OTUs for the 65 Rasborinae taxa, 2.5-fold more than by using morphological characters. The number of OTUs observed within species ranged

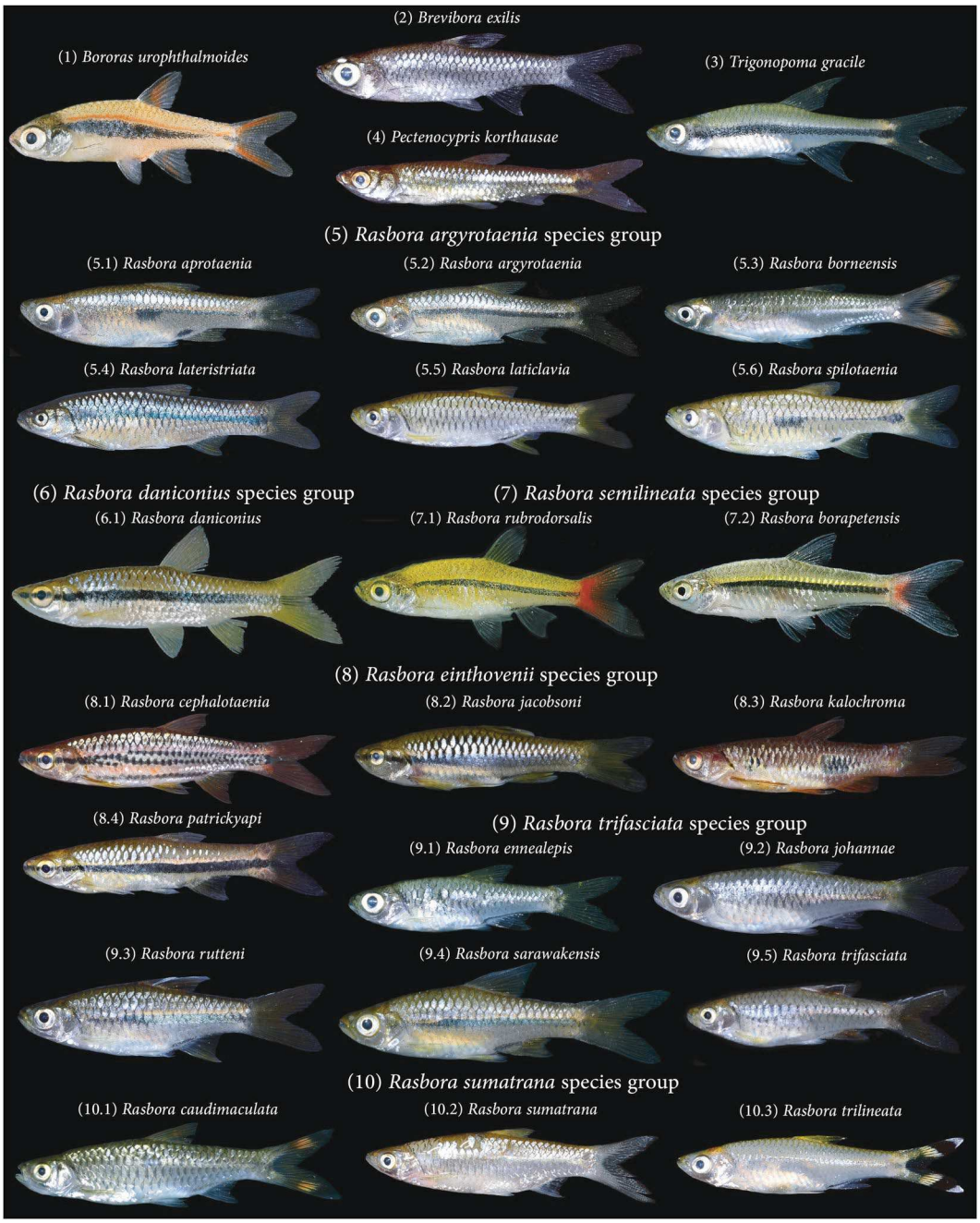

Figure 1. Selected species of Rasborinae that illustrate the diversity of the subfamily in Sundaland. All pictures, except 1, 6.1, 7.1 and 7.2, originate from the Barcode of Life Datasystem (dx.doi.org/110.5883/DS-BIFRA, Creative Commons Attribution - Non Commercial - Share Alike), pictures 1, 6.1 and 7.2 originate from FFish.asia (https://ffish.asia, Creative Commons Attribution - Non Commercial - Share Alike).

from two for 22 species to 11 for *Trigonostigma pauciperforata* (Table 1). Based on the results of the species delimitation analyses, a re-examination of the original species identity associated with 105 DNA barcodes mined from BOLD and GenBank revealed 13 cases of conflicts that likely originated from mis-identifications (Table S4). These concerned the genera *Boraras* (four records, two species), *Brevibora* (two records, two species) and *Rasbora* (seven records, six species). Along the same line, 12 uncertain identifications were revised for the genera *Rasbora* (10 records, five taxa) and *Trigonostigma* (two records, one taxa).

The examination of the maximum K2P genetic distances for species with multiple OTUs and within OTUs revealed large differences with maximum K2P distances ranging between 0.26 and 13.64 within species and between 0.00 and 2.37 within OTUs (Table 1). This trend was largely confirmed by the distribution of the genetic distances at both species and OTUs levels (Fig. 4). At the species level, the distribution of the maximum intraspecific K2P genetic distance broadly overlap with the distribution of the K2P distances to the nearest neighbor (Fig. 4A,B, Table S5) and no barcoding gap is observed. On average, the nearest neighbor K2P genetic distances

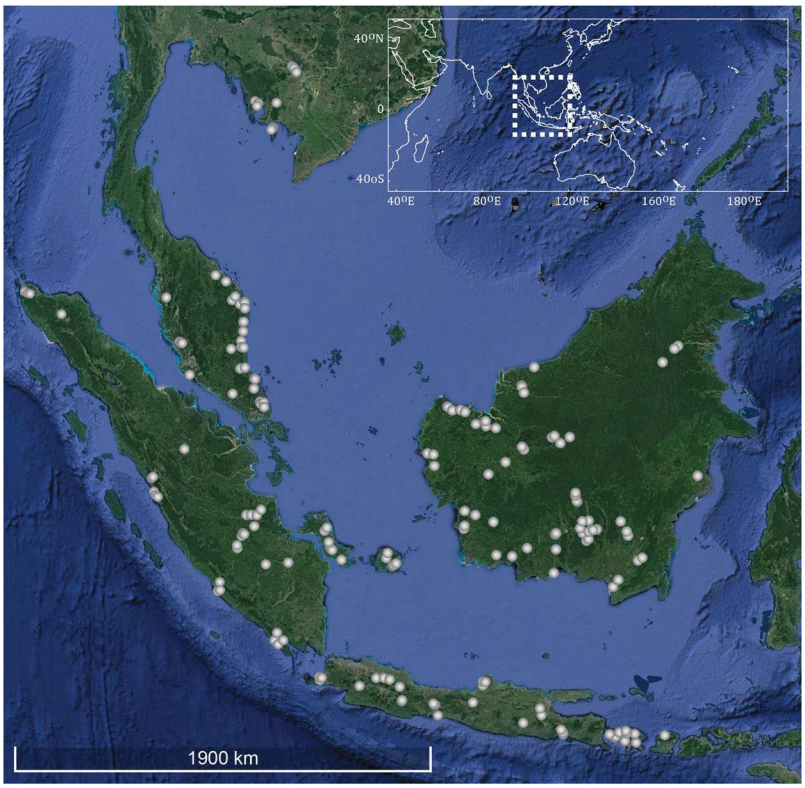

Figure 2. Collection sites for the newly generated 991 samples analyzed here. Each dot may represent several collection sites. Map data: Google, DigitalGlobe. Modified using Adobe Illustrator CS5 v 15.0.2. http://www.adobe.com/products/illustrator.html.

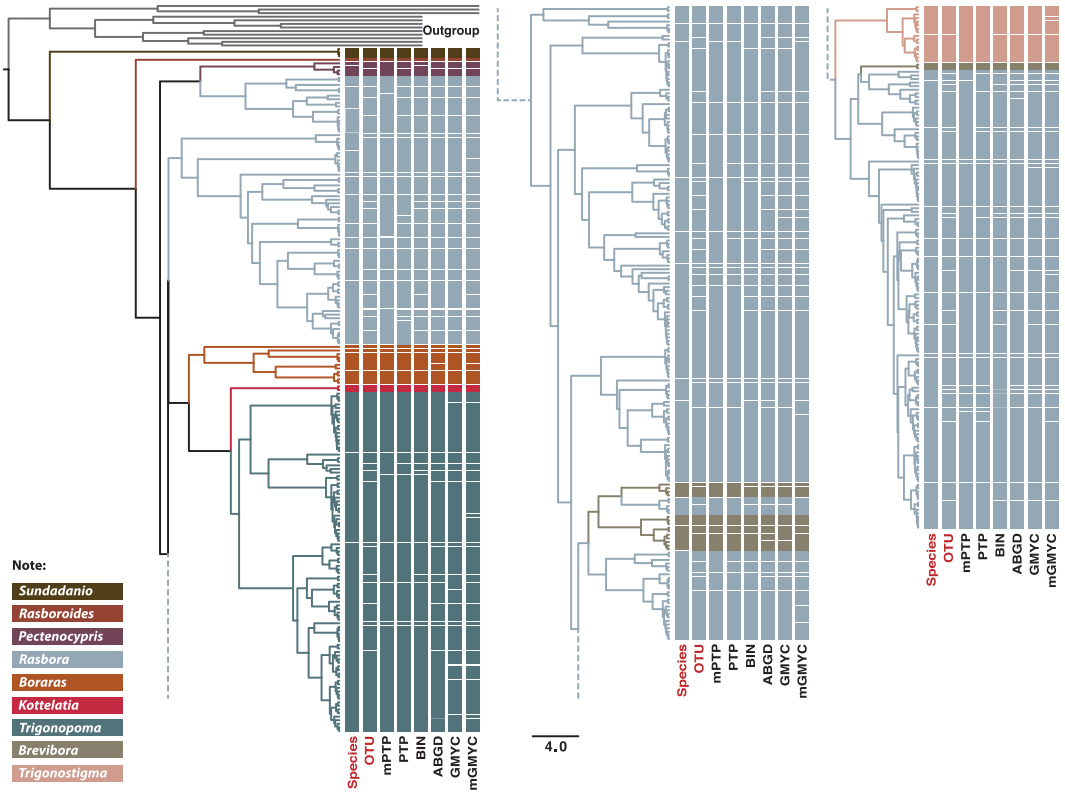

Figure 3. Bayesian maximum credibility tree of the Rasborinae DNA barcodes (identical haplotypes removed) and species delimitation according to GMYC, mGMYC, PTP, mPTP, ABGD, BIN and the 50% consensus delimitation.

are only 3.5-fold higher than the maximum intraspecific K2P distances. Plotting genetic distance for each species provides little improvement as a substantial number of species display maximum intraspecific K2P genetic distances higher than the minimum distance to the nearest neighbor (Fig. 4C). At the OTU level, the overlap is drastically reduced peaking between 0 and 0.99 for maximum intraspecific K2P distances and ranging between 1.0 and 1.99 for the K2P distance to the nearest neighbor (Fig. 4D,E). The distribution width of the maximum intraspecific K2P distance is much more restricted for OTUs than species and fewer cases of maximum intraspecific distance higher than the minimum distance to the nearest neighbor are observed (Fig. 4F). At the OTU level, the nearest neighbor K2P genetic distances were 7.2-fold higher on average than the maximum intraspecific K2P genetic distances.

Range distributions inferred from the new records generated for this study indicate that most type localities are embedded in the observed species range (Fig. 5). The degree of overlap between species range, however, largely varies among genera with little or no overlap observed for *Boraras*, *Pectenocypris*, *Trigonostigma* and most *Rasbora* species while a substantial amount of overlap is observed for *Brevibora* and *Trigonopoma* species (Fig. 5).

Discussion

This study represents the most comprehensive molecular survey conducted for the subfamily Rasborinae[19,34]. Our DNA barcode reference library consists of 65 Rasborinae species distributed across 7 genera and covering 78% of the Rasborinae diversity reported from Sundaland. DNA barcoding delivers reliable species-level identifications when taxa possess unique COI sequence clusters characterized by multiple private mutations. This condition was met for all the Rasborinae species examined here and no cases of retention of ancestral polymorphism were detected[35]. However, this clearly contrasts with multiples discrepancies observed within the set of previously published COI sequences obtained on GenBank and BOLD. About 25 percent of these records were either misidentified or associated with uncertain identifications. Such mis-identifications were expected considering the morphological uniformity within some Rasborinae genera, particularly in the genus *Rasbora* where multiple cases of taxonomic conflicts have been highlighted already[16,36–39]. Unexpectedly, most of the conflicts we detected were within the larger species of *Rasbora*, particularly those of the *Rasbora argyrotaenia* group and the *R. sumatra* group, and not within closely related smaller species such as members of the *R. trifasciata* group (Fig. 1). In facts, conflicts in species level population assignments have been previously reported for the *R. argyrotaenia* group in Java and Bali where *R. lateristriata* and *R. baliensis* have been confounded for decades as recently revealed by re-examination of species boundaries and distribution through DNA barcodes[16]. Other morphologically similar species of the *Rasbora argyrotaenia* group have been previously confused with *R. lateristriata*, such as *R. elegans*, *R. spilotaenia* and *R. chrysotaenia*. These species are difficult to separate due to overlapping meristic counts and coloration patterns[40]. Our study, however, highlights that these species have disjunct range distributions (Fig. 5) and cluster into well-differentiated mitochondrial lineages (Fig. S1, Table S3). Several of the detected misidentifications also involve species from different *Rasbora* species groups[24] such as *Rasbora dusonensis*, from the *R. argyrotaenia* group, that has been previously mistaken for *R. sumatrana* from the *sumatrana* group and *R. myersi*, from the *R. sumatrana* group, that has been confounded with *R. dusonensis* from the *argyrotaenia* group. Despite being distantly related (Fig. S1), these species show overlapping meristic counts and similar coloration patterns with no dark spots on the body[40]. This result further calls for a broader assessment of the monophyly of the different *Rasbora* groups, previously identified by Liao[24], as they are poorly supported by our study (Fig. S1).

The observed average ratio of 3.5 between intraspecific and interspecific distances is very low compared to earlier values found for the Javanese ichtyofauna, where minimum nearest neighbor genetic distances are on average 28-fold higher than the maximum intraspecific genetic distances[27]. This value is also very low in comparison to previous large-scale fish DNA barcode surveys[41–46]. This deviation can be attributed to a substantial amount of cryptic diversity revealed by our species delimitation analyses. For 61 species, delimitated on the basis of morphological characters, and validated by a match between species range distributions and type localities, we recovered a total of 166 OTUs. When accounting for this cryptic diversity the ratio between the minimum nearest neighbor and maximum intraspecific distances rose to 7.5. Earlier large scale surveys in Sundaland already pointed to substantial levels of cryptic diversity[28–31,33] and it has also been demonstrated that small-size species are more sensitive to fragmentation, experience faster genetic drift and as such accumulate cryptic diversity at a faster rate than large-size species[45,47]. Along the same line, small-size species are more frequently confounded and lumped together, a bias that tend to inflate the proportion of hidden diversity[48].

We found very high numbers of OTUs with deep genetic divergences (up to 13.64% in *Trigonopoma gracile*) in a number of species (ranging from 7 to 11) such as in *Rasbora bankanensis*, *Rasbora einthovenii*, *Rasbora trilineata*, *Trigonopoma gracile* and *Trigonopoma pauciperforatum*. These five species also display some of the widest range distributions in Sundaland with OTUs occurring in Borneo, Sumatra, Peninsular Malaysia and several small islands across the Java sea (*R. bankanensis*, Fig. 5(16); *R. einthovenii*, Fig. 5(19); *R. trilineata*, Fig. 5(8); *T. gracile*, Fig. 5(5); *T. pauciperforatum*, Fig. 5(4)). However, the scarcity of OTU range overlap for those species suggests ongoing population fragmentation across the species range distribution (Tables S2 and S3). This pattern is likely connected to the complex geological history of Sundaland which over the last 10 Million years was influenced by the subduction activity of the Asian and Australian plates and the resulting intense volcanic activity which produced multiple volcanic arches[5]. Furthermore, climatic fluctuations during the Pleistocene induced major sea levels changes leading to merging of Sundaland landmasses during glacial maxima and multiple fragmentations during glacial sea level low-stands[7,8]. In such dynamic landscapes, complex patterns of distribution and high lineage diversity are to be expected[10]. The influence of eustatic fluctuations in Sundaland is exemplified by *Rasbora bankanensis*, *Rasbora einthovenii*, *Rasbora trilineata*, *Trigonopoma gracile* and *Trigonopoma pauciperforatum* all of which display wide range distributions among watersheds neighboring the Java sea. Those have been repeatedly connected during glacial maxima (Fig. 5(5), 5(8), 5(16) and 5(19)). This pattern strongly contrasts with the lower genetic diversity and restricted range distribution of the species occurring in the Eastern part of Borneo such as *Rasbora vaillantii*

Species/OTUs	Max. Intraspecfic Dist. (%)	Nearest Neighbor Dist. (%)
Brevibora cheya	4.29	5.99
OTU 105 (BOLD:AAY0408)	0.00	3.18
OTU 106 (BOLD:ADN0681)	1.30	3.18
Brevibora dorsiocellata	2.10	7.71
OTU 102 (BOLD:ADY4509)	0.00	1.57
OTU 103 (BOLD:ADN0680)	0.52	1.57
Rasbora aprotaenia	1.83	1.04
OTU 140 (**BOLD:ADY6054**)	1.30	1.04
OUT 139 (BOLD:ADZ0447)	NA	1.30
Rasbora argyrotaenia	5.67	5.97
OTU 87 (BOLD:ADY7291)	0.26	5.10
OTU 88 (BOLD:ACQ2593)	0.52	5.10
Rasbora arundinata	2.63	2.10
OTU 130 (BOLD:ADF6073)	0.00	2.10
OTU 131 (BOLD:ADN1040)	0.00	2.63
Rasbora bankanensis	7.12	6.51
OTU 40 (BOLD:ACF1059)	GenBank	GenBank
OTU 39 (BOLD:AAR2899)	0.00	1.30
OTU 38 (BOLD:ADY4700)	NA	1.30
OTU 41 (BOLD:ADY2504)	0.00	2.91
OTU 42 (BOLD:ADY1545)	0.00	3.72
OTU 43 (BOLD:ADY1544)	0.00	2.91
OTU 44 (BOLD:ACC0430)	1.04	1.57
OTU 144 (BOLD:ADY5341)	1.04	1.57
Rasbora beauforti	2.37	9.79
OTU 33 (**BOLD:ADY4385**)	NA	2.10
OTU 34 (**BOLD:ADY4385**)	0.26	2.10
Rasbora borapetensis	7.73	5.97
OTU 86 (BOLD:ADY1548)	NA	7.43
OTU 91 (BOLD:AAU5232)	0.52	5.97
Rasbora caudimaculata	1.83	7.71
OTU 100 (BOLD:ADO5236)	0.00	1.83
OTU 101 (BOLD:AAR2916)	NA	1.83
Rasbora cephalotaenia	6.84	7.68
OTU 4 (BOLD:ADY2668)	2.36	5.41
OTU 5 (BOLD:AAI0355)	GenBank	GenBank
OTU 6 (BOLD:ADN8441)	0.26	3.17
OTU 7 (BOLD:AAI0356)	0.78	3.17
Rasbora daniconius	0.26	11.18
OTU 2 (BOLD:ABX6594)	GenBank	GenBank
OTU 3 (BOLD:ACA0514)	0.26	11.18
Rasbora dusonensis	1.57	10.73
OTU 10 (BOLD:AAU2983)	0.26	1.30
OTU 9 (BOLD:ADN2767)	0.00	1.30
Rasbora einthovenii	11.10	8.31
OTU 73 (BOLD:ADY2667)	NA	7.45
OTU 74 (BOLD:ADY1546)	0.00	7.45
OTU 75 (BOLD:ADY1017)	0.00	7.75
OTU 77 (BOLD:ADW2748)	GenBank	GenBank
OTU 78 (BOLD:ADN0813)	0.00	5.43
OTU 79 (BOLD:AAU5112)	0.00	3.18
OTU 80 (BOLD:ADO6360)	NA	2.10
OTU 81 (BOLD:ADY1549)	1.57	2.10
OTU 83 (BOLD:ADY0551)	0.52	1.30
OTU 82 (BOLD:ADY0550)	0.78	1.30
Rasbora elegans	1.57	1.04
Continued		

Species/OTUs	Max. Intraspecfic Dist. (%)	Nearest Neighbor Dist. (%)
OTU 138 (**BOLD:ADY6054**)	0.00	1.04
OTU 136 (BOLD:ADY7956)	NA	1.30
OTU 137 (BOLD:ADZ0446)	0.00	1.30
Rasbora ennealepis	9.14	6.51
OTU 35 (BOLD:ADN3883)	0.00	5.94
OTU 36 (BOLD:ADN3887)	0.78	3.97
OTU 37 (BOLD:ADY4386)	0.26	3.97
Rasbora jacobsoni	0.00	8.84
OTU 66 (BOLD:ADW4597)	GenBank	GenBank
OTU 67 (BOLD:ADN9402)	0.00	8.84
Rasbora kalbarensis	0.52	12.42
OTU 20 (BOLD:AAY0409)	GenBank	GenBank
OTU 21 (BOLD:ADN1457)	0.52	12.42
Rasbora kalochroma	2.64	5.39
OTU 71 (**BOLD:AAR2898**)	NA	1.30
OTU 72 (**BOLD:AAR2898**)	1.83	1.30
Rasbora kottelati	2.37	5.39
OTU 68 (BOLD:ADX8298)	GenBank	GenBank
OTU 69 (BOLD:ADN0290)	0.00	2.10
OTU 70 (BOLD:ADX9355)	0.26	2.10
Rasbora lateristriata	1.83	2.90
OTU 141 (BOLD:ACQ7159)	1.30	1.57
OTU 142 (BOLD:ACQ7160)	0.00	1.57
Rasbora laticlavia	4.55	4.00
OTU 119 (BOLD:ADN8626)	NA	3.45
OTU 120 (BOLD:ADO3612)	0.78	1.04
OTU 121 (BOLD:ADY6696)	0.26	1.04
Rasbora patrickyapi	1.83	8.31
OTU 146 (**BOLD:ADN2766**)	0.00	1.83
OTU 76 (**BOLD:ADN2766**)	0.00	1.57
OTU 147 (**BOLD:ADN2766**)	NA	1.57
Rasbora paucisqualis	5.97	7.63
OTU 26 (BOLD:ADY2665)	0.00	4.28
OTU 27 (BOLD:ADX9120)	0.00	2.63
OTU 28 (BOLD:ADY4316)	NA	1.57
OTU 29 (BOLD:ADY4315)	NA	1.57
OTU 30 (BOLD:ADY4317)	0.26	1.83
Rasbora paviana	2.37	2.36
OTU 126 (**BOLD:AAD6182**)	0.52	1.83
OTU 127 (**BOLD:AAD6182**)	GenBank	GenBank
OTU 129 (BOLD:ADY6053)	1.30	1.83
Rasbora rutteni	6.22	7.14
OTU 17 (BOLD:ADN4430)	1.04	5.93
OTU 18 (BOLD:ADY4516)	0.00	2.37
OTU 19 (BOLD:ADN7331)	0.00	2.37
Rasbora sp.1	2.63	3.45
OTU 124 (BOLD:ACQ2698)	0.00	2.63
OTU 125 (BOLD:ACQ2594)	0.52	2.63
Rasbora subtilis	3.99	7.06
OTU 111 (BOLD:ADN7332)	NA	3.99
OTU 112 (BOLD:ADN3888)	1.57	3.99
Rasbora sumatrana	3.18	5.97
OTU 89 (**BOLD:AAY0407**)	1.04	2.37
OTU 90 (**BOLD:AAY0407**)	0.78	2.37
Rasbora tornieri	1.57	9.51
OTU 84 (**BOLD:ADL5624**)	NA	1.57
Continued		

Species/OTUs	Max. Intraspecfic Dist. (%)	Nearest Neighbor Dist. (%)
OTU 85 (**BOLD:ADL5624**)	0.00	1.57
Rasbora trilineata	10.97	7.69
OTU 96 (BOLD:AAE7383)	1.83	2.36
OTU 97 (BOLD:ADN9095)	0.00	5.96
OTU 98 (BOLD:ADN7260)	NA	3.74
OTU 99 (BOLD:ADN9096)	0.00	3.74
OTU 93 (**BOLD:AAE7384**)	GenBank	GenBank
OTU 94 (**BOLD:AAE7384**)	0.00	4.27
OTU 95 (BOLD:ADY1696)	0.26	2.36
Rasbora tuberculata	9.87	9.73
OTU 24 (BOLD:ADN3886)	0.00	9.55
OTU 25 (BOLD:ADN3884)	0.26	9.55
Rasbora vaillantii	1.83	4.00
OTU 117 (BOLD:ADY8198)	0.00	1.83
OTU 118 (BOLD:ADY8199)	0.00	1.83
Rasbora vulcanus	0.00	7.69
OTU 115 (BOLD:AAI0352)	GenBank	GenBank
OTU 116 (BOLD:ADN3885)	0.00	7.69
Trigonopoma gracile	13.64	9.22
OTU 46 (BOLD:ADN4644)	NA	6.26
OTU 47 (BOLD:ADO0069)	0.00	1.57
OTU 48 (BOLD:ADY2669)	NA	2.36
OTU 49 (BOLD:ADY4282)	NA	1.57
OTU 50 (BOLD:ADY6176)	0.00	1.57
OTU 145 (**BOLD:ACC0899**)	0.52	2.10
OTU 51 (**BOLD:ACC0899**)	2.10	2.10
Trigonopoma pauciperforatum	9.25	9.22
OTU 55 (BOLD:ADY1547)	1.83	1.83
OTU 56 (BOLD:ADY1425)	0.00	1.83
OTU 57 (BOLD:ADY5548)	0.52	2.10
OTU 58 (BOLD:AAV7972)	NA	4.81
OTU 59 (BOLD:ACC0580)	1.30	4.54
OTU 60 (**BOLD:ADY2666**)	0.00	1.30
OTU 61 (**BOLD:ADY2666**)	1.57	1.30
OTU 62 (BOLD:ADN4643)	NA	3.45
OTU 63 (BOLD:ACC0669)	0.00	3.99
OTU 64 (BOLD:ADV1540)	2.37	1.83
OTU 65 (BOLD:AAY0427)	1.83	1.83
Trigonostigma heteromorpha	2.37	2.64
OTU 108 (BOLD:AAJ8936)	0.26	1.83
OTU 110 (BOLD:ABZ6147)	0.26	1.83

Table 1. List of the morphological species displaying more than one OTU including the maximum intraspecific and minimum nearest neighbor K2P distances for species and OTUs.

(Fig. 5(10)), *R. laticlavia* (Fig. 5(10)), *R. trifasciata* (Fig. 5(15)) and *R. reophila* (Fig. 5(20)) or species occurring in the Western part of Sumatra such as *Rasbora vulcanus* (Fig. 5(9)), *R. maninjau* (Fig. 5(9)), *R. jacobsoni* (Fig. 5(9)), *R. tawarensis* (Fig. 5(10)); *R. chrysotaenia* (Fig. 5(11)) and *R. arundinata* (Fig. 5(11)) and species in Java and Bali such as *Rasbora* sp1 (Fig. 5(10)), R. sp2 (Fig. 5(14)), *R. lateristriata* (Fig. 5(14)) and *R. baliensis* (Fig. 5(14)). These parts of Borneo, Sumatra and partially Java were disconnected from the central region of Sundaland around the Java sea during the Pleistocene. This trend highlights the sensitive status of the endemic Rasborinae species in the peripheral areas of Sundaland due to their highly restricted distribution ranges. The present study also argues against translocation programs for the most widespread species, considering the high proportion of cryptic diversity, if species and OTUs identity are not determined through DNA barcodes[16,31].

Conclusions

The subfamily Rasborinae is the most diverse freshwater fish group of Sundaland and therefore represents an excellent model to explore the evolutionary response of local freshwater biotas to a dynamic geological history and repeated eustatic fluctuations. Affected by taxonomic confusions for decades, the genus *Rasbora* has been left

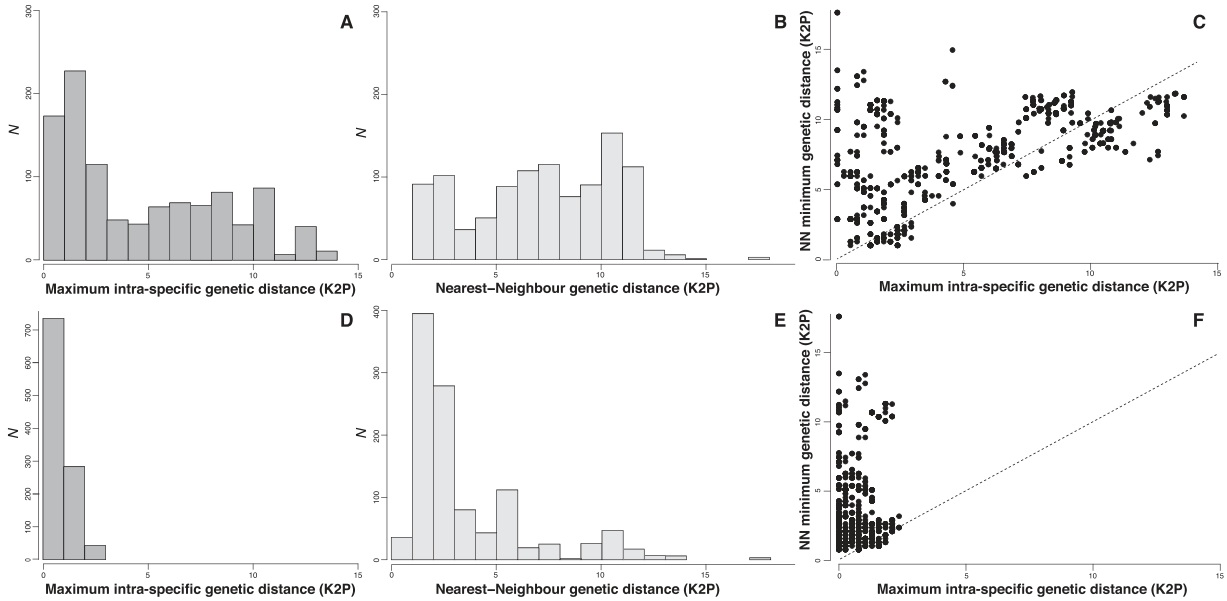

Figure 4. Summary of the distribution of the K2P distances. (**A**,**D**) Maximum intraspecific K2P distances; (**B**,**E**) Minimum nearest-neighbor K2P distances; (**C**,**F**) Individual plotting of maximum intraspecific K2P distances and minimum nearest neighbor K2P distances. (**A**–**C**) Distributions of K2P distances for species delimitated using morphological characters. (**D**–**F**) Distributions of K2P distances for OTUs delimitated by the 50% consensus among species delimitation methods.

aside of recent large-scale molecular studies aimed at exploring the diversification of aquatic biotas in Sundaland. Our comprehensive DNA barcode reference library for the subfamily enables further evolutionary studies on the diversification of the group, in particular within the genus *Rasbora*, which allowed us to trace evolutionary dynamics at the local scale in Sundaland[16]. The contrasting patterns of molecular diversity and species range distributions between Rasborinae species inhabiting the watersheds neighboring the Java sea and the species located on the Eastern part of Borneo call for a larger assessment of their dynamics of species proliferation based on broader genomic analyses. Clearly, future studies will also have to address the systematics of the Rasborinae as no evidence supporting the monophyly of *Rasbora* nor the different *Rasbora* species groups are detected here.

Material and Methods

Sampling and collection management. Material used in the present study is the result of a collective effort to assemble a global Rasborinae DNA barcode reference library through various field sampling efforts conducted by several of the coauthors in Sundaland over the past decade. Specimens were captured using gears such as electrofishing, seine nets, cast nets and gill nets across sites that encompass the diversity of freshwater lentic and lotic habitats in Sundaland (Fig. 2). Specimens were identified following original descriptions where available, as well as monographs[40,49]. Species names were further validated using several online catalogs[50,51]. Specimens were photographed, individually labeled and voucher specimens were preserved in a 5% formalin solution. Prior to fixation a fin clip or a muscle biopsy was taken and fixed separately in a 96% ethanol solution for further genetic analyses. Both tissues and voucher specimens were deposited in the national collections at the Museum Zoologicum Bogoriense (MZB), Research Center for Biology (RCB), Indonesian Institute of Sciences (LIPI).

Assembling a checklist of the Sundaland Rasborinae. A checklist of the Rasborinae species occurring in Sundaland was assembled from available online catalogs including Fishbase[51] and Eschmeyer's Catalog of Fishes[50] as detailed in Hubert *et al.*[15]. This checklist was used to estimate the taxonomic coverage of the present DNA barcoding campaign and to identify type localities for each species. The following information was included: (1) authors of the original description, (2) type locality, (3) latitude and longitude of the type locality, (4) holotype and paratypes catalog numbers, (5) distribution in Sundaland. This information is available as online Supplementary Material (Table S1).

Sequencing and international repositories. Genomic DNA was extracted using a Qiagen DNeasy 96 tissue extraction kit following manufacturer's specifications. A 651-bp segment from the 5′ region of the cytochrome oxidase I gene (COI) was amplified using primers cocktails C_FishF1t1/C_FishR1t1 including M13 tails[52]. PCR amplifications were done on a Veriti 96-well Fast (ABI-AppliedBiosystems) thermocycler with a final volume of 10.0 µl containing 5.0 µl Buffer 2× 3.3 µl ultrapure water, 1.0 µl each primer (10 µM), 0.2 µl enzyme Phire Hot Start II DNA polymerase (5U) and 0.5 µl of DNA template (~50 ng). Amplifications were conducted as followed: initial denaturation at 98 °C for 5 min followed by 30 cycles denaturation at 98 °C for 5 s, annealing at 56 °C for 20 s and extension at 72 °C for 30 s, followed by a final extension step at 72 °C for 5 min. The PCR products were purified with ExoSap-IT (USB Corporation, Cleveland, OH, USA) and sequenced in both directions. Sequencing reactions were performed using the "BigDye Terminator v3.1 Cycle Sequencing Ready Reaction" and

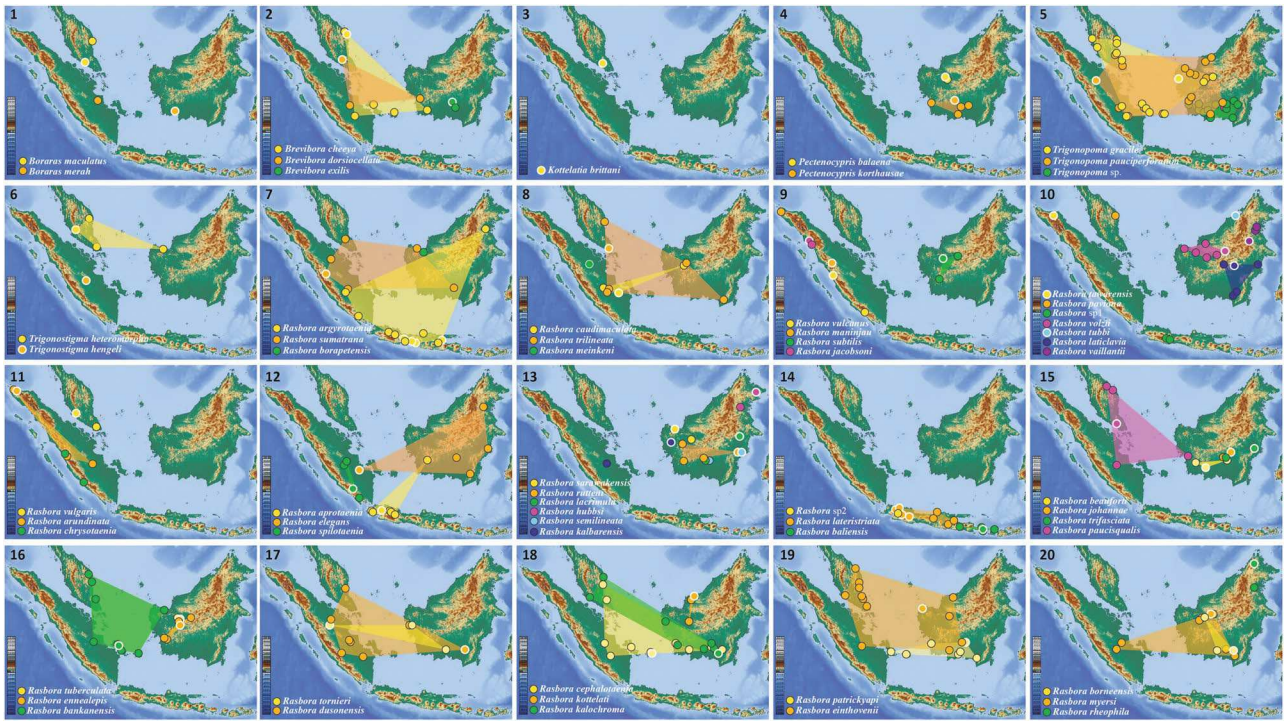

Figure 5. Maps depicting species distribution ranges as established based on the present sampling sites (black margin) and type localities (white margin) following the checklist generated for this study (Table S1). **1** Sampling sites and type localities of *Boraras maculatus* and *B. merah*. **2** Sampling sites, type localities and distribution ranges of *Brevibora cheeya*, *B. dorsiocellata* and *B. exilis*. **3** Type locality of *Kottelatia brittani*, sampling sites unknown, sequences originating from GenBank. **4** Sampling sites, type localities and distribution ranges of *Pectenocypris korthausea* and *P. balaena*. **5** Sampling sites, type localities and distribution ranges of *Trigonopoma gracile*, *T. pauciperforatum* and *T.* sp[16]. **6** Sampling sites, type localities and distribution ranges of *Trigonostigma heteromorpha*, *T. hengeli* (sampling sites outside the map); *T. espei* not displayed, type locality outside the map and sampling sites unknown, sequences originating from GenBank. **7** Sampling sites, type localities and distribution ranges of *Rasbora argyrotaenia*, *R. sumatrana* and *R. borapetensis*, multiple type localities for *Rasbora argyrotaenia* as detailed in[16], Type locality of *R. borapetensis* outside the map. **8** Sampling sites, type localities and distribution ranges of *Rasbora caudimaculata*, *R. trilineata* and *R. meinkeni*, sampling sites of *R. meinkeni* unknown, sequence originating from GenBank. **9** Sampling sites, type localities and distribution ranges of *Rasbora vulcanus*, *R. maninjau*, *R. subtilis* and *R. jacobsoni*. **10** Sampling sites, type localities and distribution ranges of *Rasbora tawarensis*, *R. paviana*, R. sp. 1[16], *R. volzii*, *R. tubbi*, *R. laticlavia* and *R. vaillanti*, sampling sites corresponding to the type locality for *Rasbora tawarensis*. **11** Sampling sites, type localities and distribution ranges of *Rasbora vulgaris*, *R. arundinata* and *R. chrysotaenia*, type locality of *Rasbora chrysotaenia* located in Sumatra with no further details (Table S1). **12** Sampling sites, type localities and distribution ranges of *Rasbora aprotaenia*, *R. elegans* and *R. spilotaenia*. **13** Sampling sites, type localities and distribution ranges of *Rasbora sarawakensis*, *R. rutteni*, *R. lacrimula*, *R. hubbsi*, *R. semilineata* and *R. kalbarensis*. **14** Sampling sites, type localities and distribution ranges of *Rasbora* sp2[16], *R. lateristriata* and *R. baliensis*. **15** Sampling sites, type localities and distribution ranges of *Rasbora beauforti*, *R. johannae*, *R. trifasciata* and *R. paucisqualis*. **16** Sampling sites, type localities and distribution ranges of *Rasbora tuberculata*, *R. ennealepis* and *R. bankanensis*. **17** Sampling sites, type localities and distribution ranges of *Rasbora tornieri* and *R. dusonensis*. **18** Sampling sites, type localities and distribution ranges of *Rasbora cephalotaenia*, *R. kottelati* and *R. kalochroma*. **19** Sampling sites, type localities and distribution ranges of *Rasbora patrickyapi* and *R. einthovenii*. **20** Sampling sites, type localities and distribution ranges of *Rasbora borneensis*, *R. myersi*, *R. rheophila*. Sampling sites and type locality of *R. daniconius* not displayed, outside the map. Each locality may represent several sampling sites. Map data: https://maps-for-free.com/. Modified using Adobe Illustrator CS5 v 15.0.2. http://www.adobe.com/products/illustrator.html.

sequencing was performed on the automatic sequencer ABI 3130 DNA Analyzer (Applied Biosystems). DNA barcodes obtained at the Naturhistorisches Museum Bern were generated as previously described in Conte-Grand *et al.*[33].

The sequences and associated information were deposited on BOLD[53] and are available in the data set DS-BIFRA (Table S2, dx.doi.org/10.5883/DS-BIFRA). DNA sequences were submitted to GenBank (accession numbers are accessible directly at the individual records in BOLD). An additional set of 106 Rasborinae COI sequences were downloaded from GenBank (Table S3).

Genetic distances and species delimitation. Kimura 2-parameter (K2P)[54] pairwise genetic distances were calculated using the R package Ape 4.1[55]. Maximum intraspecific and nearest neighbor genetic distances were calculated from the matrice of pairwise K2P genetic distances using the R package Spider 1.5[56]. We checked for the presence of a barcoding gap, *i.e.* the lack of overlap between the distributions of the maximum intraspecific and the nearest neighbor genetic distances[57], by plotting both distances and examining their relationships on an individual basis instead of comparing both distributions independently[58]. A neighbor-joining (NJ) tree was built based on K2P distances using PAUP 4.0a[59] in order to visually inspect genetic distances and DNA barcode clusters (Fig. S1). This NJ tree was rooted using *Sundadanio retiarius*.

Several alternative methods have been proposed for delimitating molecular lineages[60–63]. Each of these methods have pitfalls, particularly when it comes to singletons (*i.e.* delimitated lineages represented by a single sequence) and a combination of different approaches is increasingly used to overcome potential pitfalls arising from uneven sampling[16,43,64–66]. We used four different sequence-based methods of species delimitation. For the sake of clarity, we refer to species identified based on morphological characters as species while species delimited using DNA sequences are referred to as Operational Taxonomic Unit (OTU)[67–69]. OTUs were delimitated using the following algorithms: (1) Refined Single Linkage (RESL) as implemented in BOLD and used to generate Barcode Index Numbers (BIN)[62], (2) Automatic Barcode Gap Discovery (ABGD)[61], (3) Poisson Tree Process (PTP) in its multiple rates version (mPTP) as implemented in the stand-alone software mptp_0.2.3[63,70], (4) General Mixed Yule-Coalescent (GMYC) in its multiple rate version (mGMYC) as implemented in the R package Splits 1.0–19[71]. RESL and ABGD used DNA alignments as input files while a ML tree was used for mPTP and a Bayesian Chronogram based on a strict-clock model using a 1.2% of genetic distance per million year[72] for mGMYC. The mPTP algorithm uses a phylogenetic tree as an input file, thus, a maximum likelihood (ML) tree was first reconstructed using RAxML[73] based on a GTR + Γ substitution model. Then, an ultrametric and fully resolved tree was reconstructed using the Bayesian approach implemented in BEAST 2.4.8[74]. Two Markov chains of 50 millions each were ran independently using the Yule pure birth model tree prior, a strict-clock model and a GTR + I + Γ substitution model. Trees were sampled every 10,000 states after an initial burnin period of 10 millions. Both runs were combined using LogCombiner 2.4.8 and the maximum credibility tree was constructed using TreeAnnotator 2.4.7[74]. Identical haplotypes were pruned for further species delimitation analyses.

Received: 21 November 2019; Accepted: 20 January 2020;
Published online: 18 February 2020

References

1. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
2. Lamoreux, J. F. *et al.* Global tests of biodiversity concordance and the importance of endemism. *Nature* **440**, 212–214 (2006).
3. Hoffman, M. *et al.* The impact of Conservation on the status of the world's vertebrates. *Science (80-.)* **330**, 1503–1509 (2010).
4. Schipper, J. *et al.* The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science (80-.)* **322**, 225–230 (2008).
5. Lohman, K. *et al.* Biogeography of the Indo-Australian archipelago. *Annu. Rev. Ecol. Evol. Syst.* **42**, 205–226 (2011).
6. Hall, R. Late Jurassic-Cenozoic reconstructions of the Indonesian region and the Indian ocean. *Tectonophysics* **570–571**, 1–41 (2012).
7. Woodruff, D. S. Biogeography and conservation in Southeast asia: how 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugium-phase biodiversity. *Biodivers. Conserv.* **19**, 919–941 (2010).
8. Voris, H. K. Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *J. Biogeogr.* **27**, 1153–1167 (2000).
9. De Bruyn, M. *et al.* Borneo and Indochina are major evolutionary hotspots for Southeast Asian biodiversity. *Syst. Biol.* **63**, 879–901 (2014).
10. De Bruyn, M. *et al.* Paleo-drainage basin connectivity predicts evolutionary relationships across three Southeast Asian biodiversity hotspots. *Syst. Biol* **62**, 398–410 (2013).
11. O'Connell, K. A. *et al.* Within-island diversification underlies parachuting frog (*Rhacophorus*) species accumulation on the Sunda shelf. *J. Biogeogr.* **45**, 929–940 (2018).
12. O'Connell, K. A. *et al.* Diversification of bent-toed geckos (*Cyrtodactylus*) on Sumatra and west Java. *Mol. Phylogenet. Evol.* **134**, 1–11 (2019).
13. Hendriks, K. P., Alciatore, G., Schilthuizen, M. & Etienne, R. S. Phylogeography of Bornean land snails suggests long-distance dispersal as a cause of endemism. *J. Biogeogr.* (2019).
14. Dong, J. *et al.* Biogeographic patterns and diversification dynamics of the genus *Cardiodactylus* Saussure (Orthoptera, Grylloidea, Eneopterinae) in Southeast Asia. *Mol. Phylogenet. Evol.* **129**, 1–14 (2018).
15. Hubert, N. *et al.* DNA barcoding Indonesian freshwater fishes: challenges and prospects. *DNA Barcodes* **3**, 144–169 (2015).
16. Hubert, N. *et al.* Revisiting species boundaries and distribution ranges of *Nemacheilus* spp. (Cypriniformes: Nemacheildae) and *Rasbora* spp. (Cypriniformes: Cyprinidae) in Java, Bali and Lombok through DNA barcodes: implications for conservation in a biodiversity hotspot. *Conserv. Genet.* **20**, 517–529 (2019).
17. Keith, P. *et al. Schismatogobius* (Gobiidae) from Indonesia, with description of four new species. *Cybium* **41**, 195–211 (2017).
18. Conway, K. W., Hirt, M. V, Yang, L., Mayden, R. L. & Simons, A. M. Conway, K. W., Hirt, M. V., Yang, L., Mayden, R. L., & Simons, A. M. Cypriniformes: Systematics & Paleontology: Festschrift in honor of G. Arratia. In *Origin and Phylogenetic Interrelationships of Teleosts* 295–316 (2010).
19. Tang, K. *et al.* Systematics of the subfamily Danioniae (Teleostei: Cypriniformes: Cyprinidae). *Mol. Phylogenet. Evol.* **57**, 189–214 (2010).
20. Stout, C. C., Tan, M., Lemmon, A. R., Lemmon, E. M. & Armbruster, J. W. Resolving Cypriniformes relationships using an anchored enrichment approach. *BMC Evol. Biol.* **16**, 244 (2016).
21. Hirt, M. V. *et al.* Effects of gene choice, base composition and rate heterogeneity on inference and estimates of divergence times in cypriniform fishes. *Biol. J. Linn. Soc* **121**, 319–339 (2017).
22. Tan, M. & Armbruster, J. W. Phylogenetic classification of extant genera of fishes of the order Cypriniformes (Teleostei: Ostariophysi). *Zootaxa* **4476**, 6–39 (2018).
23. Brittan, M. R. A revision of the Indo-Malayan frash-water fish genus *Rasbora*. *Monogr. Inst. Sci. Tech. Manila* **3**, 3–pls (1954).

24. Liao, T. Y., Kullander, S. O. & Fang, F. Phylogenetic analysis of the genus *Rasbora* (Teleostei: Cyprinidae). *Zool. Scr* **39**, 155–176 (2010).
25. Kottelat, M. & Vidthayanon, C. *Boraras micros*, a new genus and species of minute freshwater fish from Thailand (Teleostei: Cyprinidae). *Ichthyol. Explor. Freshwaters* **4**, 161–176 (1993).
26. Kottelat, M. & Witte, K.-E. Two new species of *Microrasbora* from Thailand and Myanmar, with two new generic names for small Southeast Asian cyprinid fishes (Teleostei: Cyprinidae). *J. South Asian Nat. Hist* **4**, 49–56 (1999).
27. Dahruddin, H. *et al*. Revisiting the ichthyodiversity of Java and Bali through DNA barcodes: Taxonomic coverage, identification accuracy, cryptic diversity and identification of exotic species. *Mol. Ecol. Resour.* **17**, 288–299 (2017).
28. Nurul Farhana, S. *et al*. Exploring hidden diversity in Southeast Asia's *Dermogenys* spp. (Beloniformes: Zenarchopteridae) through DNA barcoding. *Sci. Rep* **8**, 10787 (2018).
29. Beck, S. *et al*. Plio-Pleistocene phylogeography of the Southeast Asian Blue Panchax killifish, *Aplocheilus panchax*. *PLoSONE* **12**, e0179557 (2017).
30. Lim, H.-C., Abidin, M. Z., Pulungan, C. P., De Bruyn, M. & Mohd Nor, S. A. DNA barcoding reveals high cryptic diversity of freshwater halfbeak genus *Hemirhamphodon* from Sundaland. *PLoSONE* **11**, e0163596 (2016).
31. Hutama, A. *et al*. Identifying spatially concordant evolutionary significant units across multiple species through DNA barcodes: Application to the conservation genetics of the freshwater fishes of Java and Bali. *Glob. Ecol. Conserv* **12**, 170–187 (2017).
32. Nguyen, T. T. T., Na-Nakorn, U., Sukmanomon, S. & ZiMing, C. A study on phylogeny and biogeography of mahseer species (Pisces: Cyprinidae) using sequences of three mitochondrial DNA gene regions. *Mol. Phylogenet. Evol.* **48**, 1223–1231 (2008).
33. Conte-Grand, C. *et al*. Barcoding snakeheads (Teleostei, Channidae) revisited: Discovering greater species diversity and resolving perpetuated taxonomic confusions. *PLoS One* **12**, e0184017 (2017).
34. Collins, R. A. *et al*. Barcoding and border biosecurity: identifying cyprinid fishes in the aquarium trade. *PLoSONE* **7**, e28381 (2012).
35. Funk, D. J. & Omland, K. E. Species-level paraphyly and polyphyly: frequency, causes and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Syst.* **34**, 397–423 (2003).
36. Siebert, D. J. The identities of *Rasbora paucisqualis* Ahl in Schreitmuller, 1935, and *Rasbora bankanensis* (Bleeker, 1853), with the designation of a lectotype for R. paucisqualis (Teleostei: Cyprinidae). *Raffles Bull. Zool.* **45**, 29–37 (1997).
37. Kottelat, M. *Rasbora rheophila*, a new species of fish from northern Borneo(Teleostei: Cyprinidae). *Rev. Suisse Zool* **119**, 77–87 (2012).
38. Ng, H. H. & Kottelat, M. The identity of the cyprinid fishes *Rasbora dusonensis* and *R. tornieri* (Teleostei: Cyprinidae). *Zootaxa* **3635**, 62–70 (2013).
39. Muchlisin, Z. A., Fadli, N. & Siti-Azizah, M. N. Genetic variation and taxonomy of *Rasbora* group (Cyprinidae) from Lake Laut Tawar, Indonesia. *J. Ichthyol* **52**, 284–290 (2012).
40. Kottelat, M., Whitten, A. J., Kartikasari, S. R. & Wirjoatmodjo, S. *Freshwater Fishes of Western Indonesia and Sulawesi* . (Periplus editions, 1993).
41. Hubert, N. *et al*. Identifying Canadian freshwater fishes through DNA barcodes. *PLoS One* **3**, e2490 (2008).
42. April, J., Mayden, L. R., Hanner, R. H. & Bernatchez, L. Genetic calibration of species diversity among North America's freshwater fishes. *Proc. Natl. Acad. Sci. USA* **108**, 10602–10607 (2011).
43. Shen, Y. *et al*. DNA barcoding the ichthyofauna of the Yangtze River: insights from the molecular inventory of a mega-diverse temperate fauna. *Mol. Ecol. Resour.* **19**, 1278–1291 (2019).
44. Hubert, N. *et al*. Cryptic diversity in indo-pacific coral-reef fishes revealed by DNA-barcoding provides new support to the centre-of-overlap hypothesis. *PLoS One* **7**, e28987 (2012).
45. Hubert, N. *et al*. Geography and life history traits account for the accumulation of cryptic diversity among Indo-West Pacific coral reef fishes. *Mar. Ecol. Prog. Ser.* **583**, 179–193 (2017).
46. Pereira, L. H. G., Hanner, R., Foresti, F. & Oliveira, C. Can DNA barcoding accurately discriminate megadiverse Neotropical freshwater fish fauna? *BMC Genet.* **14**, 20 (2013).
47. April, J., Hanner, R., Mayden, R. L. & Bernatchez, L. Metabolic rate and climatic fluctuations shape continental wide pattern of genetic divergence and biodiversity in fishes. *PLoSONE* **8**, e70296 (2013).
48. Kottelat, M., Britz, R., Tan, H. H. & Witte, K.-E. *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. *Proc. R. Soc. London, B* **273**, 895–899 (2006).
49. Kottelat, M. The fishes of the inland waters of Southeast Asia: A catalog and core bibliography of the fishes known to occur in freshwaters, mangroves and estuaries. *Raffles Bull. Zool* **Supplement 27**, 1–663 (2013).
50. Eschmeyer, W. N., Fricke, R. & van der Laan, R. Catalog of fishes electronic version. (2018).
51. Froese, R. & Pauly, D. FishBase. Available at, http://www.fishbase.org (2014).
52. Ivanova, N. V., Zemlak, T. S., Hanner, R. H. & Hebert, P. D. N. Universal primers cocktails for fish DNA barcoding. *Mol. Ecol. Notes* **7**, 544–548 (2007).
53. Ratnasingham, S. & Hebert, P. D. N. BOLD: The Barcode of Life Data System, www.barcodinglife.org. *Mol. Ecol. Notes* **7**, 355–364 (2007).
54. Kimura, M. A Simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
55. Paradis, E., Claude, J. & Strimmer, K. E. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
56. Brown, S. D. J. *et al*. Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Mol. Ecol. Resour* **12**, 562–565 (2012).
57. Meyer, C. & Paulay, G. DNA barcoding: Error rates based on comprehensive sampling. *Plos* **3**, 2229–2238 (2005).
58. Blagoev, G. A. *et al*. Untangling taxonomy: A DNA barcode reference library for Canadian spiders. *Mol. Ecol. Resour.* **16**, 325–341 (2015).
59. Swofford, D. L. Version 4.0 b10. PAUP*. Phylogenetic Anal. Using Parsimony (*Other Methods) (2001).
60. Pons, J. *et al*. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol* **55**, 595–606 (2006).
61. Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Mol. Ecol* **21**, 1864–1877 (2012).
62. Ratnasingham, S. & Hebert, P. D. N. A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoSONE* **8**, e66213 (2013).
63. Zhang, J., Kapli, P., Pavlidis, P. & Stamatakis, A. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* **29**, 2869–2876 (2013).
64. Kekkonen, M. & Hebert, P. D. N. DNA barcode-based delineation of putative species: Efficient start for taxonomic workflows. *Mol. Ecol. Resour.* **14**, 706–715 (2014).
65. Kekkonen, M., Mutanen, M., Kaila, L., Nieminen, M. & Hebert, P. D. N. Delineating species with DNA Barcodes: A case of taxon dependent method performance in moths. *PLoS One* **10**, e0122481 (2015).
66. Blair, C. & Bryson, J. R. W. Cryptic diversity and discordance in single-locus species delimitation methods within horned lizards (Phrynosomatidae: *Phrynosoma*). *Mol. Ecol. Resour.* **17**, 1168–1182 (2017).
67. Avise, J. C. Molecular Markers, Natural History and Evolution. (1989).
68. Moritz, C. Defining 'Evolutionary significant units' for conservation. *Trends Ecol. Evol.* **9**, 373–375 (1994).

69. Vogler, A. P. & DeSalle, R. Diagnosing units of conservation management. *Conserv. Biol.* **6**, 170–178 (1994).
70. Kapli, P. *et al*. Multi-rate Poisson Tree Processes for single-locus species delimitation under Maximum Likelihood and Markov Chain Monte Carlo. *Bioinformatics* **33**, 1630–1638 (2017).
71. Fujisawa, T. & Barraclough, T. G. Delimiting species using single-locus data and the generalized mixed Yule coalescent approach: A revised method and evaluation on simulated data sets. *Syst. Biol* **62**, 707–724 (2013).
72. Bermingham, E., McCafferty, S. & Martin, A. P. Fish biogeography and molecular clocks: Perspectives from the Panamanian Isthmus. In *Molecular Systematics of* Fishes (eds. Kocher, T. D. & Stepien, C. A.) 113–128 (CA Academic Press, 1997).
73. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
74. Bouckaert, R. R. *et al*. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).

Acknowledgements

The authors wish to thank Siti Nuramaliati Prijono, Bambang Sunarko, Witjaksono, Mohammad Irham, Marlina Adriyani, Ruliyana Susanti, Rosichon Ubaidillah, the late Renny K. Hadiaty, Hari Sutrisno and Cahyo Rahmadi at Research Centre for Biology (RCB-LIPI) in Indonesia; Edmond Dounias, Jean-Paul Toutain, Robert Arfi and Valérie Verdier from the 'Institut de Recherche pour le Développement'; Joel Le Bail and Nicolas Gascoin at the French embassy in Jakarta for their continuous support. We also would like to thank Eleanor Adamson, Hendry Budianto, Tob Chann Aun, Pak Epang, Herman Ganatpathy, Sébastien Lavoué, Michael Lo, Hendry Michael, Joshua Siow, Heok Hui Tan, Elango Velautham, Norsham S. Yaakob, and Denis Yong for their help in the field. We are also particularly thankful to Sumanta at IRD Jakarta for his help during the field sampling in Indonesia. Part of the present study was funded by the Institut de Recherche pour le Développement (UMR226 ISE-M and IRD through incentive funds) to N.H.), the MNHN (UMR BOREA) to P.K., the French Ichthyological Society (SFI) to P.K., the Foundation de France to P.K., the French embassy in Jakarta to N.H., the Natural Environmental Research Council (NERC, NE/F003749/1) to L.R. and Ralf Britz; National Geographic (8509-08) to L.R. and North of England Zoological Society-Chester Zoo to L.R. The present study and all associated methods were carried out in accordance with relevant guidelines and regulation of the Indonesian Ministry of Research and Technology (Indonesia), the Economic Planning Unit, Prime Minister's Department (Malaysia), the Forest Department Sarawak (Malaysia), the Vietnam National Museum of Nature (Vietnam) and the Inland Fisheries Research and Development Institute (Cambodia). Field sampling in Indonesia was conducted according to the research permits 097/SIP/FRP/SM/IV/2014 for Philippe Keith, 60/EXT/SIP/FRP/SM/XI/2014 for Frédéric Busson, 41/EXT/SIP/FRP/SM/VIII/2014 for Nicolas Hubert, 200/E5/E5.4/SIP/2019 for Erwan Delrieu-Trottin and, 1/TKPIPA/FRP/ SM/I/2011 and 3/TKPIPA/FRP/SM/III/2012 for Lukas Rüber. The Fieldwork in Peninsular Malaysia and Sarawak was conducted under permits issued by the Economic Planning Unit, Prime Minister's Department, Malaysia (UPE 40/200/19/2417 and UPE 40/200/19/2534) and the Forest Department Sarawak (NCCD.970.4.4[V]-43) and were obtained with the help of Norsham S. Yaakob (Forest Research Institute Malaysia, Kepong, Kuala Lumpur, Malaysia). Luong Van Hao and Pham Van Luc (Vietnam National Museum of Nature) helped with arranging research permits in Vietnam and So Nam (Inland Fisheries Research and Development Institute, IFReDI) helped with arranging research permits in Cambodia. All experimental protocols were approved by the Indonesian Ministry of Research and Technology (Indonesia), the Indonesian Institute of Sciences (Indonesia), the Forest Department Sarawak (Malaysia), Economic Planning Unit of the Prime Minister's Department (Malaysia), the Vietnam National Museum of Nature (Vietnam) and the Inland Fisheries Research and Development Institute (Cambodia). It is a great pleasure to thank Soraya Villalba for generating the DNA barcodes at the Naturhistorisches Museum Bern. Sequence analysis was aided by funding through the Canada First Research Excellence Fund as part of the University of Guelph Food from Thought program. We thank Paul Hebert, Alex Borisenko and Evgeny Zakharov as well as BOLD and CCDB staff at the Centre for Biodiversity Genomics, University of Guelph for their valuable support. This publication has the ISEM number 2019-293-SUD.

Author contributions

L.R. and N.H. designed the study. A.S., T.S., H.D., R.R., R.E., A.W., K.K., F.B., S.S., U.N., E.D., I.V.U., Z.A.M., D.W., P.K., L.R. and N.H. conducted the field sampling. A.S., E.D.T., T.S., H.D., A.W., L.R. and N.H. performed the morphological identifications. H.D., S.S., U.N., F.B. and N.H. curated the specimen collection. A.S., E.D.T., H.D., M.S.A.Z., Y.F., I.V.U., R.H., D.S., L.R. and N.H. conducted the laboratory work. A.S., E.D.T., H.D., F.B., N.H., R.H. and D.S. curated the DNA barcode records in BOLD. A.S., E.D.T., H.D., J.F.A., L.R. and N.H. analyzed the data. A.S., E.D.T., H.D., J.F.A., D.S., L.R. and N.H. wrote the initial manuscript and all authors commented and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-59544-9.

Correspondence and requests for materials should be addressed to N.H.

Reprints and permissions information is available at www.nature.com/reprints.

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 Author(s) 2020

Acknowledgements

I thank Allah SWT for the opportunities, both to live my life and learn so many things about “Life” by conducting this study. “*Verily, in the Heavens and the Earth are signs for the believers. And in your creation, and what He scattered (through the Earth) of moving (living) creatures are signs for people who have Faith with certainty.*” (Al-Jathiyah (45:3-4). While indeed, all of this is about understanding the distribution pattern of organism on Earth and the mechanism behind that, honestly though, it’s not easy for me to understand that by reading and studying Scriptures only. Thus, it has been a pleasure to eventually start to know more about things by doing Science.

Aside of God Almighty and His never ending blesses, this study will also never be completed, not even started, without helps, endorsements and encouragements from so many people whom I will mention in my following "scribbles" or leave them out when it's better to keep them only in my heart.

I shall thank my family for always offering a comforting environment for me to develop sense of curiosity in Science. Even though, most of them is actually working on Faith by profession. Which is perfectly strange and beautiful in many senses. My parents (Nina Nurjanah and Asep Ruhendi) and my little sisters (Arni Muslimah Handayani Widjaja and Arni Puji Fajriyah Hadi Widjaja) who are always somehow understanding and forgiving for having a kid and a big sister who's rather "particular" and hard to deal with like me. (Alm.) Ne Uum and Engki Fandi Affandi; (alm.) Pak Ustadz E. Syamsudin and Ma Cicih; (alm.) Aa A. Tjakra Widjaja & (alm.) Omah Naromas; (alm.) Ki Ukat and their extended families.

I shall thank my teachers for their contributions in encouraging and supporting my study in the field of Phylogeography, Evolution and Biodiversity. Pak Djoko T. Iskandar for introducing and opening up a whole different world to me, who never fail to inspire me to be a better Biologist. Pak Achmad Sjarmidi (Pak Mamid) for always being the exceptional teacher in many aspects. For being patient and generous to me all these years since the time I was kicked out (by him) from my seat in his class, until now when I, usually, just throw him random chats via WhatsApp, day and night. Bu Tati S. Syamsudin, Bu Pingkan Aditiawati and Bu Endah Sulistyawati for the challenge

and trust that they have, as well as for their supports since I was preparing for my study and scholarship, till in the end when I could finally wrap them up.

I shall thank my thesis supervisor Jean-François Agnèse and co-supervisor Nicolas Hubert for trusting and guiding me through this study despite I had no particular research background in this field. Some people start things from zero, but I think I started all of this from minus. Especially, I thank Nico for coming to Bandung by himself to talk about this opportunity as well as for bridging and accomodating my stay in Indonesian Institute of Sciences - Cibinong during the first year of this study. As much as this was a “journey” for me, I believe that it was a total roller coaster for them. I am really grateful for their patient and dedication despite all of my shortcomings. Thank you for not giving up on me.

I thank *Institut des Sciences de l’Évolution de Montpellier* (ISEM) and GAIA Doctoral School for hosting me during my study in France. I thank members and previous members of *Evolution des Poissons* (EPOISS) research team from ISEM for their cooperation and supports: Patrick Berrebi, Bruno Guinand, Jean-François Baroiller, Helena Dcota, Antoine Pariselle, Christelle Tougard, Cecile Triay, Juliette Pouzadoux, Hala Ainou, Madoka Krick, Pierre Caminade. Especially, I am very grateful to be able to work in a “bubble” with Erwan Delrieu-Trottin, Hadi Dahruddin and Sélim Ben Chéhida. Erwan has always been my reference for many things, especially for technological crash courses. Not to mention, I really thank him for saving my computer, and all of my data, nearing the end of my study. GAIA Doctoral School has always been supportive and helpful, especially in issuing related documents for the administrative requirements of my scholarship. For that, I thank Marc Bouvy, Cedrine Jay-Allemand and Aida Dubost.

I thank Pak Hari Sutrisno and Pak Cahyo Rahmadi as well as staffs of Zoology division (Research Centre for Biology-Indonesian Institute of Sciences, Cibinong) for hosting me during my first year. Especially, I also thank Ibu Daisy Wowor who has been very supportive for my research both during my stay in Cibinong as well as after I started to stay in France and write my articles, and Pak Ujang Nurhaman for helping me during field samplings. I thank Edmond Dounias, Frédéric Busson and Pak Sumanta from the *Institut de Recherche pour le Développement* (IRD); as well as Joel Le Bail and Nicolas Gascoin at the French embassy in Jakarta for their continuous support. I thank Paul Hebert, Alex Borisenko and Evgeny Zakharov as well as BOLD

and CCDB staffs at the Centre for Biodiversity Genomics, University of Guelph for their valuable supports during sequence analyses for this study.

I thank Fabien L. Condamine, Lukas Rüber, Laurent Pouyaud, Philippe Keith, Pak Tedjo Sukmono, Sopian Sauri, Renny Risdawati, Roza Elvyra, Pak Arif Wibowo, Bu Kustiati, Pak Muhamad Syamsul Arifin Zein, Yuli Fitriana, Ilham Vemendra Utama, Zainal Abidin Muchlisin, Robert Hanner, Dirk Steinke and Marie-Ka Tilak as co-authors of my articles, both published and in press. I am particularly grateful for Lukas Rüber and Laurent Pouyaud who have been sharing their molecular datasets for my study as well as Fabien Condamine for his significant contributions in the diversification analyses. I thank all members of my *comité de suivi de thèse*: Sophie Arnaud-Haond, Patrick Berrebi (previously from EPOISS), Philippe Borsa, Christelle Tougard (previously from EPOISS) and Christelle Hely. I thank Nicolas Puillandre and Mark de Bruyn who are willing to be *rapporteurs* of my thesis defense alongside with Philippe Keith and Emmanuel Douzery as *examinateurs,* as well as Fabien Condamine and Lukas Rüber as *invités*. I also thank David J. Lohman for his relevant comments and encouragement on one of my articles.

I thank *Lembaga Pengelola Dana Pendidikan* (LPDP) – *Kementerian Keuangan* RI, for their financial support by awarding me *Beasiswa Pendidikan Indonesia* (BPI). In particular, I thank Ibu Ratna Prabandari who offered crucial decision so that I can start this study properly as well as Pak Dwi Larso for issuing revision of my Letter of Guarantee (LoG) thus I can wrap up my study in peace. I also offered my belated greeting to (Alm.) Pak Surna Tjahja Djajadiningrat (Pak Naya), a founding father of School of Business and Management ITB. I thank him for taking a very indirect yet very important part on “me getting the scholarship” thus I could start all of this.

I thank Madame Laure Guigou from CIHEAM-IAMM for facilitating me with accomodation during my study. I especially grateful that CIHEAM-IAMM modified their regulation this year as a response to the special condition resulted from Covid-19 pandemic, thus I needed not to leave the dorm through out the summer break. Furthermore, I thank the institution to manage the dorm really well even in the middle of this pandemic so I can still feel as safe as ever. I also thank Prof. Didier Bessis (Dermatology Departement – Hospital Saint-Eloi), Dr. Clement Boissin (Pneumology Departement – Hospital Arnaud de Villeneuve) of CHU-Montpellier as well as all of the medical staffs that have been, literally, saving my life. This physical limitation is both annoying and somehow dangerous. I thank all of them for helping me to “stay afloat”,

both by medically helping to manage my health as well as to give more understanding so that I can actually taking a better care of my own self.

I thank *Perhimpunan Pelajar Indonesia* (PPI) Montpellier, its members and previous members, as well as *Ibu-Ibu* Franco-Indo in Montpellier for being great friends during my stay in the city. It was nice to feel not “being alone” during my study. In particular, I am grateful to get to know Nurul Novelia Fuandila during my last year. I thank her also for letting me to befriend Sarthak Malusare. Both of you really painted the end of my study with rainbows. The colours resisted all of the lockdown and gloomy things happening this year.

Doing this study means a lot for me as it paves my scientific experiences to be continued and developed in the future while working as fellow faculty member of my institution. I shall not forget my teachers and colleagues for their never ending supports; for being inspirational by being “their own selves”; for making such environment that ensured me since long ago to live my life with “Life Sciences” as enjoyable and as colourful as they do. For that, I thank Bu Devi N. Choesin, Bu Dea I. Astuti, Pak Agus D. Permana, Bu Rina Ratnasih, Pak Gede Suantika, Teh Dian Rosleine, Teh Dzulianur Mutsla, Kak Neil Priharto, Teh Maya Fitriyanti, Teh Novi Tri Astutiningsih, Kang Donny K. Hardjani, Kak Fenryco Pratama, Aditya D. Pramudya, members of Ecology research group as well as all teaching and administrative staffs of School of Life Sciences and Technology ITB.

Lastly, I shall share my extensive gratefulness for people who have always been there. My comrades in many crimes. Vilandri Astarini, a soon-to-be mother of two. Who has been my friend since my first days in ITB. I’m really grateful that she can still offer me a comfort of friendship despite she is currently living a very different life. Noviana Budianti, a fellow PhD student/deadliner from University of Shizuoka. Similar backgrounds grow similar thinking organisms, us. I hope your thesis defense will go smoothly, and let’s make our way to Subang later on to harvest our long waited natural bounties in Vilandri’s (well, her father’s) orchards.

Rahman Rasyidi, my fellow “*Pejuang*” who has been friend in high and low, always “synced-in” despite many differences. Who had been blown up together by below freezing nocturnal winds of Netherland, stormed by merciless tropical rains in Bandung, fried by dominating sun of Pangandaran and Batu Karas. Friend in managing countless classes, practicums and field trips. Friend in the struggling of power, such retaliation towards the mainstreams. Ganjar Cahyadi, the real young generation

Taxonomist. Someone who is both knowledgable and naive, decisive and humble. An exemplary scientist that has been managing the “Treasury” for a while, while I’m gone. I have always been sorry for leaving you alone with such hard works all these years. But then again, I know that you are happy as it is also your passion, so you will forgive me anyway. Even, there is a big chance that you would think that I should not be sorry after all. For both of you who have been sharing vision with me, we all know what we should do, what we will do, because we have already had the list. Thank you for being patient. We will start it right away.

In harmonia progressio…
Demi Tuhan, untuk Bangsa dan Almamater!

Montpellier, 14 December 2020,

Arni SHOLIHAH

To be continued…