UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO BIOLOGIA ANIMAL

"Genomics of Speciation in Humpback Dolphins (Genus *Sousa*)"

Cátia Sofia Formas Chanfana

Mestrado em Biologia Evolutiva e do Desenvolvimento

Dissertação orientada por:
Doutora Ana Rita Amaral

2018

Dedicated to my family.

Agradecimentos | Acknowledgements

This work would not be possible without the help of amazing people beside me. It is a pleasure to thank those who were by my side through all of this endeavor.

First of all, I would like to express my profound gratitude to Doctor Ana Rita Amaral, my supervisor who was always available to support me through all the steps of this thesis. Even with all the blunders that life put in front of you, you were always concerned, supportive and gentle with me. Thank you for offering this thesis to me, it has been a pleasure to work with you. I would also like to thank Dr. Howard Rosenbaum and the Wildlife Conservation Society for the opportunity to work with their humpback dolphin dataset and thus collaborate with them in their long term project.

I want to give a special thanks for Vitor Sousa. Even though you did not have the obligation to help me, you offer teachings, patience and guidance when I needed. Thank you for everything.

For all of my colleagues in the Evolutionary Genetics group, especially for João Carvalho, Mónica Silva, Rodrigo, Teresa Santos and Ana Catarina Silva, I want to thank you all for the interesting discussions, for helping me gain some insights about my career in the future, for criticizing constructively my work and for all the patience and teaching. Thank you.

I would like to offer my thanks to all of my colleagues from FCUL, especially to Sofia Mendes and João Moreno. For all the encouragement and insightful discussions (which were invaluable), but also for providing joyous moments during this year. Without it probably this road would be much more difficult to get through. Thank you.

To João Frazão, even though you are new to the group, I would like to thank you for all the joyous conversations and sharing experiences during lunch for this few months. It has been a pleasure.

I also want to give a sincere gratitude for my both friends, André Silva and Pedro Albuquerque. Not only for the constructive critics of my thesis that both of you gave, which I am extremely grateful, but also for the joyous moments that you giving me during our late conversations, and escapes during dinner. Thank you so much.

Of course, I want to give a special thanks to my family. This journey (not only for this thesis, but also for the past six years) would not have been possible without their support and encouragement. A minha gratidão mais profunda vai para os meus pais, porque sem eles nada teria sido possível. Por tudo o que fizeram por mim, por me deixarem sonhar e não desistir, pelo amor e pela dedicação um muito obrigado. E claro ao meu irmão, pelas palhaçadas e pelos mimos que te pedia, estavas lá sempre. Obrigada do fundo do coração. I am truly blessed for having you all as my family.

And last but not the least, a huge thank you for the only person who actually endured all the good and bad moods that occur during this year. Without your help, support, patience (a lot of patience!!) and late-night discussions, this road would be extremely painful. For everything that you gave me, I hope that I am doing the same for you, but for now I just want to show my enormous gratitude for helping me, for your kindness, for being who you are. Thank you Carlos Ramírez!

Resumo

"Como se originam novas espécies?" é uma questão fundamental em Evolução e a importância de compreender os mecanismos e processos por detrás dessa origem está ligado à formação da biodiversidade. Designado por especiação, é um processo contínuo e complexo que envolve múltiplas barreiras e interações entre as mesmas que levam ao isolamento de populações e por consequente a formação de novas espécies. Estas múltiplas barreiras podem ser de dois tipos de isolamento – pré-zigótico e pós-zigótico – e cada uma destas barreiras deixa assinaturas diferentes no genoma. Com o avanço das tecnologias de *next generation sequencing*, *scans* genómicos têm vindo a ganhar extrema relevância neste campo de investigação. O rápido decréscimo do custo de sequenciação em gerar milhares de marcadores genéticos e com o desenvolvimento de novos programas que lidam com dados genómicos, veio tornar possível aos investigadores a obtenção de um número vasto de loci/genes e identificar assinaturas e padrões de heterogeneidade em diferentes espécies. Com estes novos meios torna-se possível ter uma melhor compreensão dos mecanismos genéticos que estão envolvidos na estabilização do isolamento das populações, e que leva à origem de novas espécies.

A especiação tem sido muito estudada, mas atualmente tudo o que se sabe sobre os mecanismos por detrás deste processo contínuo, deriva sobretudo de estudos realizados em espécies terrestres e de água-doce. A terra e os oceanos têm diferentes características e diferentes tipos de barreiras ao fluxo genético entre populações. Relativamente aos oceanos, barreiras como correntes oceânicas, *upwelling*, batimetria, temperatura de superfície e salinidade têm sido propostas como alguns dos fatores que explicam a diversidade genética observada em espécies marinhas, incluindo os mamíferos marinhos. Os mamíferos marinhos estão divididos em quatro diferentes grupos, e todos eles representam uma das transições evolutivas mais impressionantes entre o ambiente terrestre para o ambiente aquático. Os cetáceos, um dos grupos de mamíferos marinhos, é composto por espécies com elevada capacidade de dispersão, aparentemente sem barreiras à sua dispersão. Todavia, estudos recentes têm vindo a demonstrar que barreiras oceanográficas, comportamento e estrutura social são fatores que explicam os padrões observados de diversidade e estrutura genética nestes animais.

O género *Sousa*, pertencente à Família Delphinidae, encontra-se distribuído descontinuamente ao longo da costa Oeste Africana até à costa Este do Oceano Pacífico, e atualmente são quatro as espécies diferentes: *S. teuszii*, *S. plumbea*, *S. chinensis e S. sahulensis*. As quatro espécies são morfologicamente distintas, estando as diferenças focadas principalmente na coloração e na forma da barbatana dorsal e a corcunda existente por baixo da barbatana dorsal. Poucos estudos têm sido feitos para compreender o seu comportamento, ecologia e genética. Contudo, tem sido apontado que barreiras oceanográficas podem estar por detrás da aparente regionalização das populações. Mais recentemente, estudos genéticos demonstraram que existe uma população altamente diferenciada no Bangladesh com aparente distribuição ao longo da Baia de Bengala, e apresenta características mistas entre duas das espécies, *S. chinensis e S. plumbea*. Ao nível do DNA mitocondrial sabe-se que esta população não se agrupa com nenhuma dessas espécies, mas sim que se encontra mais próxima filogeneticamente de *S. sahulensis* que ocorre na Austrália. Com a possibilidade de reformulação da taxonomia do género *Sousa* e com a necessidade de implementar programas de conservação, quais foram os processos evolutivos que levaram a esta diversidade de populações tem sido alvo de discussão, das quais as barreiras oceanográficas têm sido apontadas como causais. Assim, deste modo, neste estudo aplicamos técnicas de *scans* genómico para estudar a complexidade do processo de especiação dentro do género *Sousa*. De modo a termos uma ideia da estrutura populacional e dos efeitos das barreiras sobre o genoma que levam à especiação nestes organismos, utilizamos a técnica de *genotyping-by-sequencing* para obter *single nucleotide polymorphisms*, dos quais observou-se padrões de variação e diferenciação genómica ao

longo da distribuição destas espécies, e possíveis assinaturas de seleção e *loci* candidatos que aparentam ter um papel no processo de especiação.

Com o objetivo de estudar a estrutura populacional do género *Sousa*, 36 amostras foram recolhidas ao longo de toda a distribuição do género abrangendo todas as espécies atualmente conhecidas, juntamente com a população do Bangladesh. Desses 36 indivíduos, devido à má qualidade de algumas das amostras apenas 32 foram usados para o *data set* final, focando a análise apenas na distribuição do Indo-Pacífico. Todas as análises realizadas ao nível da estrutura populacional apontam para que o género *Sousa* ao longo do Indo-Pacífico é composto por 5 grupos: *S. sahulensis*, *S. chinensis*, a população do Bangladesh e *S. plumbea* que está segregado em dois grupos, a da costa Africana e a do mar da Arábia. Todas os grupos surgem como altamente diferenciados uns dos outros, com exceção dos grupos de *S. plumbea* que apresentam algum fluxo genético entre elas. Esta estrutura separada em 5 grupos apresenta valores de FST elevados quando comparados com valores obtidos em comparações entre espécies de golfinhos, sendo esta estrutura também suportada por antigos trabalhos nos quais foram utilizados menos marcadores que este presente estudo. A população do Bangladesh apresenta-se altamente diferenciada das restantes, embora morfologicamente apresente características mistas de *S. plumbea* e *S. chinensis*. Voltou-se a verificar que esta população não se agrupa com nenhuma das duas espécies, e está filogeneticamente mais próxima de a *S. sahulensis*. Contudo esta população não pode ser classificada como uma nova espécie devida à falta de amostras ao longo da distribuição de *S. chinensis*. Portanto, em trabalhos futuros, para resolver a taxonomia deste género é importante incluir amostras ao longo da distribuição de *S. chinensis*.

Apesar das razões ecológicas e sociais para a explicação da elevada estruturação deste género ainda serem desconhecidas, o nosso estudo permitiu criar hipóteses diferentes das que têm sido apresentadas até à data. Neste trabalho foram evidenciados cerca de 24 genes com relevância funcional, dos quais apresentaram sinais de seleção direcional. Embora não tenha sido possível obter vias metabólicas selecionadas devido aos poucos genes usados, as descrições destes 24 genes apontam para elevadas expressões no cérebro e em tecidos do sistema reprodutor em humanos. Relativamente aos genes expressos no cérebro, todos eles apresentaram grande importância em funções neurológicas como o stress, memória, aprendizagem e circuitos emocionais. Alguns deles como os genes DRD2 e GRM7 que são recetores para diferentes neurotransmissores e que tem vindo a demonstrar importância em muitas doenças que afetam o foro social em humanos, como Esquizofrenia e Défice de atenção e hiperatividade. Já para os genes relacionados com tecidos reprodutores, embora alguns aparentem ser importantes para a formação do espermatozóide ou para a manutenção do desenvolvimento embrionário, nenhum deles foi estudado sobre o seu efeito como uma barreira pós-zigótica com implicações no isolamento das populações. No entanto o seu papel funcional pode implicar a possível formação dessa barreira.

Sabe-se que diferenças em condições ambientais influenciam a divergência entre populações, e condições oceanográficas não são diferentes. As condições extraordinárias encontradas na Baia de Bengala, como a água pouco profunda, enorme intrusão de água doce e sedimentos devido a sistemas de mangal e a um grande sistema estuarino (dos maiores do mundo), *upwellings* e reversão da corrente como mesoeddies, muito provavelmente explicam a distinção genética que se observa neste local. Contudo, com a análise dos genes candidatos existe a possibilidade que caracteres sociais estejam também a influenciar a divergência das populações no género *Sousa*. Os cetáceos são conhecidos por terem variações no seu comportamento e sistemas sociais complexos, e cada vez mais existem estudos que comprovam que as estruturas sociais afetam a divergência entre populações. Infelizmente, nos golfinhos do género *Sousa* poucos estudos de comportamento têm sido feitos em diferentes populações, e as suas associações têm sido descritas como uma estrutura de *fission-fusion*. Com pouca informação destas populações poucas causas podem ser apontadas para a explicação dos genes candidatos observados. Porém, o facto de as populações aparecerem regionalmente separadas, nós supomos que a

estruturação deste género poderá estar a ser afetada por caracteres sociais e fatores ambientais, em que ambos permitem com que as populações se mantenham isoladas geneticamente.

O esclarecimento da estrutura populacional do género *Sousa* e a compreensão dos mecanismos que levam à sua diversidade e divergência, não são só importantes para a ciência como também são de extrema importância para a conservação destas espécies. Este estudo demonstra que diversos fatores ambientais e sociais são importantes para a manutenção das populações como unidades evolutivas únicas, e que é preciso ter em conta todos estes mecanismos para auxiliar na criação de novas políticas de conservação adequadas a estas espécies, inclusivamente na criação de Áreas Marinhas Protegidas em regiões do mundo pouco desenvolvidas, mas de grande valor ecológico.

Palavras-chave: Diferenciação Genómica, Seleção, Especiação, Golfinhos-corcunda

Abstract

Speciation is a fundamental process in evolution and is important for the formation of biodiversity. It is a continuous and complex process which involves multiple interacting barriers that lead to heterogenous genomic landscapes with various peaks of divergence between populations. With the advances in next generation sequencing technologies, genome-scans became extremely important tools for this research field, due to their higher ability to obtain thousands of genetic markers. This high-density of genetic markers, along with the emergence of new analytical approaches for this type of data, made it possible to help clarify not only our understanding of the genomic basis and the evolution of genetic barriers, but also helping to unify research on both the ecological and non-ecological causes of speciation.

In this study, we applied genome-scans to gain insights on the speciation process occurring in the genus *Sousa*, not only to understand the population structure but also to find signatures of selection and possible candidate loci that may have a putative role in the establishment of divergence and speciation. Through population structure analysis we found 5 distinct clusters, clearly separating the three already known species, *S. plumbea*, *S. chinensis* and *S. sahulensis*. A slightly segregation was observed within *S. plumbea*, separating African Coast and Arabian Sea populations. The population from Bangladesh appears highly-differentiated from all others populations, supporting previous studies conducted with mtDNA.

With this highly structured genus we found possible evidence for genic divergence with putative functional relevance. From the 16 SNPs (Single Nucleotide Polymorphisms) that showed signs of directional selection, the corresponding genes are highly expressed in human tissues – brain and reproductive system – and appear to have important roles on socio-biological traits. Even though it has been hypothesized that this genus may be geographically structured due to the influence of oceanographic variables, our work shows a possible additional influence of social drivers in the maintenance of these highly isolated populations within this genus.

Keywords: Genomic Differentiation, Selection, Speciation, Humpback Dolphins

Table of Contents

Figure List

Table List

1. Introduction

Ever since Darwin, one of the most impressive facts about nature is that it is discontinuous. All animals and plants are separated in very discrete clusters, and although there is variation among individuals within a cluster, these remain discrete morphologically and genetically to each other (Coyne, 2010; Turelli et al., 2001). These clusters have been defined as species and how come these species split has been a foundational question to the field of evolutionary biology (Wolf and Ellegren, 2017). However, not all the clusters are well defined and with methodological problems that arise from this delimitation, the definition of species is one of the most discussed topics (Schwartz and Boness, 2017).

Nowadays, the definition of species is influenced by the concept, the tools used to evaluate that concept and by scientific experience (Schwartz and Boness, 2017). There are several species concepts, and all of them are distinguished by the difference between species. The Biological Species Concept (BSC) is one of the most known concepts and is defined by "a group of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups" (Mayr, 1942). The reproduction is the focus of this concept, in which species are reproductively isolated from one another by allopatric, behavioral or physiological mechanisms (such as reproductive incompatibilities) that prevent gene flow between them (Mendez et al., 2013b). The Phylogenetic Species Concept (PSC) is another concept that focuses on the ancestry of a species, suggesting that it "…is a diagnosable cluster of individuals within which there is a parental pattern of ancestry and descent, beyond which there is not, and which exhibits a pattern of phylogenetic ancestry and descent among units of like kind." (Eldredge and Cracraft, 1980). This concept implies that species are characterized by their evolutionary distinctiveness resulting from significant divergence and are usually assessed with phylogenetic method (Mendez et al., 2013b). There are other species concepts, such as the ecological species concept, the genotypic cluster, the evolutionary species concept (Mallet, 1995; Simpson, 1951; Valen, 1976), and also the population concepts that are similar to species concepts but related to ecological, evolutionary and statistical paradigms (Waples and Gaggiotti, 2006).

Even though each concept has a reason behind it, they also show limitations such the separation into species despite the ongoing interbreeding, gene flow and hybridization. Besides, some concepts like the PSC, can also artificially create new species if species distributions are very fragmented and each fragment becomes fixed for different DNA polymorphisms through the neutral process of genetic drift and not through local adaptation (Schwartz and Boness, 2017). Even with these limitations BSC remains the gold standard, most of it due to the conceptualization of Speciation at the individual level (Wu, 2001). In 2001 Wu came with the idea of the genic view of species, where species divergence occurs along a continuum of genetic differentiation, with incipient species passing through a phase where they are only partly reproductively isolated, which means that species boundaries are semipermeable. Even though this idea has long been recognized, the different concepts are still debatable in the scientific community.

The delimitation of species is usually defined using the two most knowable concepts, BSC and PSC. The reason for these two concepts is that BSC is the most known and usable concept, while PSC is less restrictive, more applicable in practice and more objective (Agapow et al., 2004). However, in this thesis we are going to use the same concepts as in previous studies of these animals, also integrating the genic view to minimize the limitations caused by the other concepts.

A process called Speciation

Speciation is the study of how new species arise. The importance of understanding the mechanisms and processes behind this origin is linked to the formation of biodiversity (Coyne, 2010). It is an often complex and continuous process that involves multiple and interacting barriers. Until it is complete the effects of this process vary along the genome and can lead to a heterogenous genomic landscape with various peaks of differentiation and divergence between populations (Ravinet et al., 2017).

Currently, Speciation is defined "as the origin of reproductive barriers among populations that permit the maintenance of genetic and phenotypic distinctiveness of these populations in geographical proximity" (Seehausen et al., 2014). These reproductive barriers can be divided in three types of isolation: the prezygotic that includes isolation through habitat, phenological or sexual; and the postzygotic, which can be separated in two forms, the extrinsic form that results from divergent or disruptive selection; and the intrinsic form, which is due to genetic incompatibilities (Feder et al., 2012; Seehausen et al., 2014). The evolution of genetic incompatibilities is independent of the environment, and the mechanisms behind these genomic conflicts have been largely studied in evolutionary biology (Wolf and Ellegren, 2017). However, recent population genomic studies of divergence across the genomes have investigated cases of ecological speciation. They have focused on extrinsic isolation and the importance of these mechanisms throughout the genome (Seehausen et al., 2014).

Different evolutionary mechanisms give rise to different genomic signatures. When speciation is driven by intrinsic barriers it often results from epistatic incompatibilities, which may accumulate either as a by-product of selection or as a result of genetic drift. Extrinsic postzygotic and prezygotic barriers may accumulate later, which facilitates both ecological coexistence between sibling species and reinforcement of reproductive isolation (Orr and Turelli, 2001; Seehausen et al., 2014). By contrast, when speciation is driven by divergent ecological or sexual selection, extrinsic postzygotic and prezygotic barriers often evolve first and interact to produce reproductive isolation, while intrinsic postzygotic barriers will evolve later during the speciation process (Marques et al., 2017). With this pattern, multiple regions are likely to be divergent and scattered across the genome. There is even theoretical arguments and empirical evidence that sites under selection in the genome will be spatially clustered when adaptive evolution proceeds under divergent selection, with either migration or recurrent hybridization (Ellegren et al., 2012; Langerhans and Riesch, 2013; Scordato et al., 2014). Regions of reduced recombination, and the accumulation of prezygotic isolation loci, may also play a role over the genomic architecture.

With all of these different signatures affecting the genome architecture, it is important to distinguish these signatures from the background pattern of the genome, to than be able to have a glimpse of the populations' history and which episodes caused the divergence between two populations (W. Wolf and Ellegren, 2016).

Speciation genomics

The central task of speciation genetics is to reconstruct the sequence in which these different barriers and factors originated in order to distinguish between causes and consequences of speciation. To achieve this, it would be ideal to take an unbiased view of the entire genome at all stages of the same process (Seehausen et al., 2014; Wolf and Ellegren, 2017). However, speciation can rarely be studied in real time in natural populations. Estimations of gene flow and the amount of variation among the loci could help determine the order in which reproductive barriers emerged, but it is challenging to make such inferences, and current methods are not accurate enough for this purpose. Nevertheless, by the integration of multiple case studies of closely related taxa that vary in their extent of divergence, inferences can be made about the time and the importance of different factors involved (Mavárez et al.,

2006). These studies have made important contributions to speciation research, and this approach is being adopted in the next generation sequencing (NGS) based genome scan.

Speciation genomics is a relatively new field that has already begun to make an important contribution to speciation research. It uses empirical data from NGS, along with the emergence of new analytical approaches for this kind of data, and it has been helped to clarify our understanding of the genomic basis, and the evolution of reproductive barriers, to unify research on both the ecological and non-ecological causes of speciation (Etter et al., 2012). NGS techniques such as Genotyping-by-sequencing (GBS) are especially appealing to use in these studies, due to the rapidly decreasing cost of high-throughput sequencing generating hundreds or thousands of neutral markers and the development of downstream genomic tools that allowed most researchers to clearly identify patterns of heterogeneity and outlier loci for large genome species (Cammen et al., 2016; Elshire et al., 2011; Narum et al., 2013). It generates large numbers of Single Nucleotide Polymorphisms (SNPs) by reducing genome complexity with restriction enzymes, which can then be used in subsequent genetic analyses and genotyping, without requiring previous genomic information (He et al., 2014; Narum et al., 2013).

Several NGS-based genome scans of population divergence have found surprisingly variable patterns of genomic divergence. The first generation of NGS-based population genomic studies of ecological speciation has therefore shown that ecological selection can cause strong isolation of small genomic regions between diverging populations and that, when reproductive isolation is strong enough to permit persistence of incipient species in sympatry, many unlinked regions typically experience significant isolation (Haasl and Payseur, 2016; Marques et al., 2017). Indeed, genome scans have shown strong isolation at genomic loci that were known to be under divergent selection. However, as already mentioned above, caution is warranted because different evolutionary processes can leave similar signatures in the genome.

Heterogeneous genomic divergence is sometimes also observed between allopatric populations of the same species in the absence of any current gene flow. Indeed, many studies assume ongoing gene flow between species, even though stochastic variation, due to recent coalescence times and incomplete lineage sorting, can lead to low divergence and high heterogeneity in a similar way, particularly when they are combined with selection (Feder et al., 2012; Meyer et al., 2016). Statistical methods are available to distinguish divergence in isolation from divergence with gene flow, and these methods are increasingly being applied to genome-scale data sets (Wolf and Ellegren, 2017). By combining multiple methods it should be possible to obtain a richer catalog of the affected loci and a better understanding of the processes involved in speciation (Chen et al., 2010).

Marine speciation and Marine Mammals

The majority of what is currently known about the patterns and processes of speciation comes from studies of terrestrial or freshwater species, where barriers to dispersal are easily observable. Land and oceans have very different features, such as density, viscosity, temperature, solubility and diffusion of oxygen differ dramatically between water and air, affecting the dispersal and distribution of marine organisms (Miglietta et al., 2011). Moreover, land and sea differ significantly in the type and effectiveness of natural barriers, which have long been considered to be much rarer in the sea than on land. This last factor has an especially important impact on an organism's potential for dispersal, thus affecting both population connectivity and speciation processes (Miglietta et al., 2011; Momigliano et al., 2017). Even though barriers have long been thought to be less common in marine ecosystems, complex oceanographic systems and land barriers such as the Isthmus of Panama and the Eastern Pacific Barrier, are known to prevent gene flow between neighbouring populations of marine taxa, originating scenarios of vicariance and in some cases of speciation (Miglietta et al., 2011). However, variables such

as ocean currents, upwelling, bathymetry, sea surface temperature, primary productivity and salinity have been proposed as some of the factors that explain genetic diversity and structure in marine organisms, including marine mammals (Amaral et al., 2017).

Marine mammals are classified into four different groups, the cetaceans (whales, dolphins and porpoises), pinnipeds (seals, sea lions and walruses), sirenians (manatees and dugongs) and marine fissipeds (polar bears and sea otters), and all of them are non-model organisms that represents one of the most striking evolutionary transitions from terrestrial to marine environments (Cammen et al., 2016; Gatesy et al., 2013). All of these groups evolved to thrive in the marine or freshwater ecosystems, not in one single occasion but in multiple independent scenarios (McGowen, 2011). Like the diversification of cetaceans and sirenians in the Eocene from the Cetartiodactyla and Afrotheria, respectively, and for pinnipeds the diversification comes around the Miocene from within the Carnivora (Foote et al., 2015; Gatesy et al., 2013). Even though the aquatic mode life required a whole set of adaptations with anatomical rearrangements, most of the phenotypic adaptations are share between the different groups (Foote et al., 2015). This makes them an exemplary system for investigating the convergent evolution of different morphological and physiological adaptations, including: the loss and reduction of many typical mammalian characteristics, such as sight and smell; and gain of other characteristics, such as echolocation, deep diving, osmoregulation and cognition (Cammen et al., 2016).

Cetaceans is a clade that is divided in two subclades, the Mysticeti (baleen whales) and Odontoceti (toothed whales), and all of them are highly mobile species with no obvious physical geographic barriers to dispersal, in comparison to the terrestrial environment (Attard et al., 2018; Cammen et al., 2016). The family Delphinidae is one of the Odontocete lineages, and has experienced an explosive radiation during the last 11 million years (McGowen, 2011). They show a wide range of ecological and behavioral diversity, but show patterns of gene flow and genetic structure that varies extensively across space and time (Bowen et al., 2016).

The recent evolutionary success of different traits of delphinids with a large-scale ocean restructuring and temperature fluctuations in the Late Miocene and Early Pliocene, have been proposed as the explanation for this radiation (Steeman et al., 2009). Some delphinid species have ecologically and behaviorally distinct groups ("ecotypes") with limited gene flow, even in parapatry or sympatry, such as killer whales that have sympatric ecotypes that differ in prey type, foraging strategy, social structure, and movement (Bowen et al., 2016). While others show strong fidelity to narrow ranges that result in genetically divergent populations along continuous coastlines or between adjacent islands, which is the case of spinner dolphins, Hector's dolphins, and Indo-Pacific humpback dolphins. In this particularly case, biogeographic boundaries such as ocean currents, salinity and temperature gradient, sea floor topology, upwelling, primary productivity and other geographic features have been proposed as some of the factors that explain their pattern of genetic diversity and structure (Amaral et al., 2017; Farhadi et al., 2017).

Genus *Sousa*

The scientific literature is rich in details about the biology and ecology of the bottlenose dolphins (*Tursiops spp.*), common dolphins (*Delphinus spp.*) and of species within the genus *Stenella*, including information of abundance, distribution, behavior, life history parameters, taxonomy and phylogenetics. However, for the humpback dolphins of the genus *Sousa* there is little information available. Until recently, very little research had been conducted and limited biological material had been obtained from these species. This difference in information is due to the widely distribution the species in question are, and to the low level of development of many of the countries surrounding their habitats, which leads to a lack of research about the genus *Sousa*. The genus *Sousa* is part of the Delphinidae family, but the

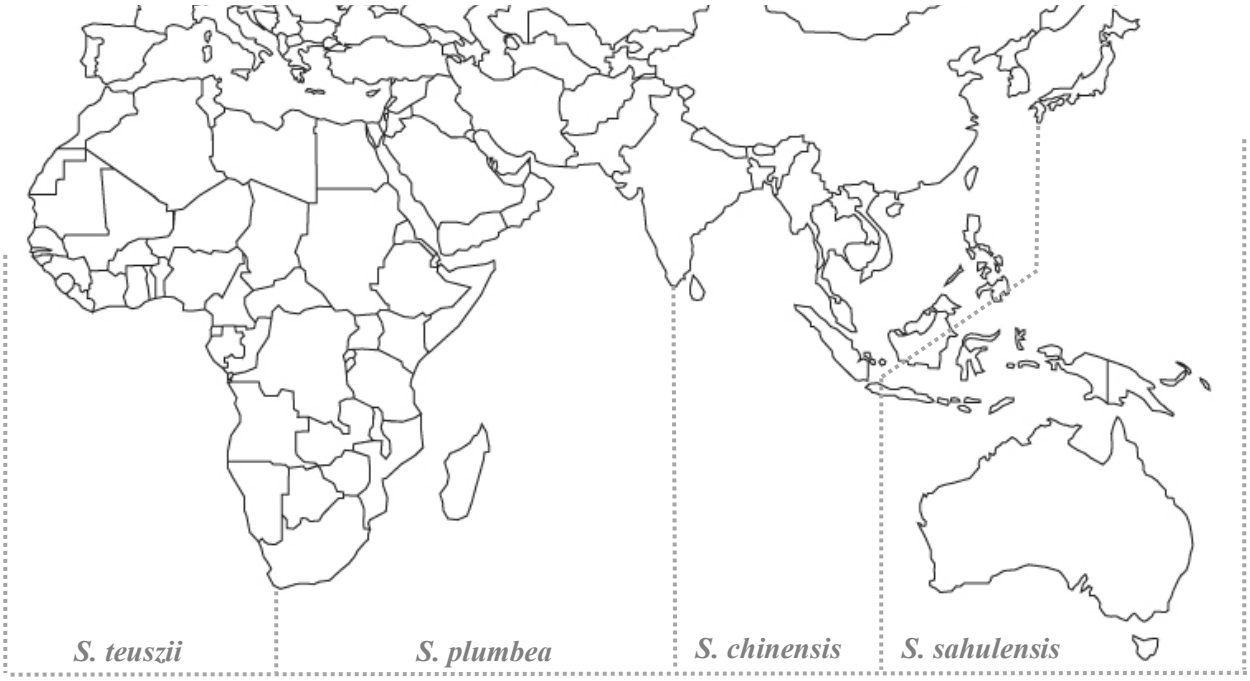

Figure 1.1 – Representation of the genus *Sousa* through their entire range. Distributed discontinuously along the coastal waters, these species appear to not move more than a few kilometers upstream, remaining in the range of tidal influence.

taxonomy within this genus has been highly controversial up until the last few years. A revision conducted in 2014 (Jefferson and Rosenbaum, 2014) presented a thorough review of morphological, molecular and biogeographic information, and suggested that this genus comprises four species: the Atlantic humpback (*S. teuszii* – East Atlantic Ocean), the Indian Ocean humpback (*S. plumbea* – Indian Ocean), Indo-Pacific humpback (*S. chinensis* – East Indian and Western Pacific Ocean), and a newly described species, the Australian humpback (*S. sahulensis* – Australia and New Guinea) as can be seen in Figure 1.1. These dolphins are distributed discontinuously along coastal waters, they occur in tropical to warm temperature regions. They tend to be in open coasts and bays, which they have access to rocky reefs, mangrove swamps, estuarine areas and areas with sandbanks or mudbanks. They do not inhabit deep oceanic areas and their movements appear to be limited by water depth (40 meters appear to be the limit) (Würsig et al., 2018). Little is known about their ecology and behavior, but it is known that these species are opportunistic-generalist feeders, eating a wide variety of coastal fishes and mostly seen in relatively small schools of less than ten individuals. Indo-Pacific humpback dolphins sometimes enter rivers and inland waterways of mangrove forests, but they do not appear to move more than a few kilometers upstream and

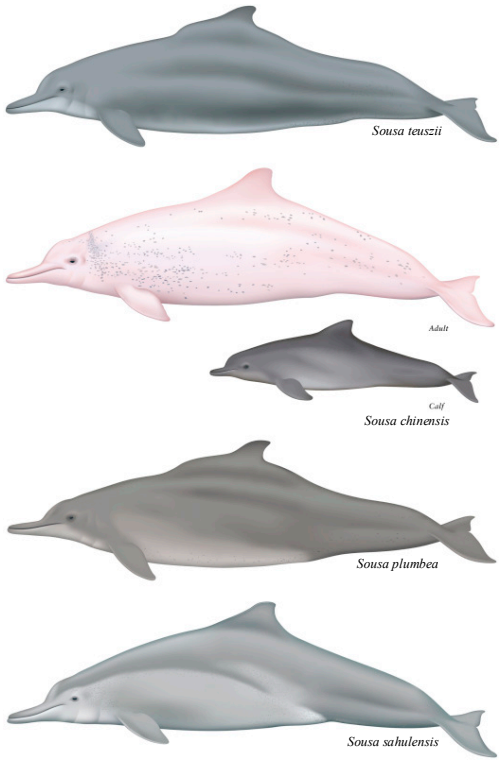

Figure 1.2 – The four recognized species of the genus *Sousa*. The Atlantic humpback dolphin is at the top, followed by the Indo-Pacific humpback dolphin adult and calf, next is the Indian Ocean humpback and the Australian humpback dolphin at the bottom. Illustration by Uko Gorter adapted from Würsig et al., 2018.

usually remain within the range of tidal influence (Jefferson and Karczmarski, 2001; Wang et al., 2007). With so much similarity over the foraging and habitat use between species, what actually distinguishes them is their visual appearance that is concentrated in three traits: pigmentation, size of the dorsal fin and the hump, which can also be regionally separated as can be seen in figure 1.2 (Amaral et al., 2017; Jefferson and Curry, 2015; Jefferson and Rosenbaum, 2014; 2016). Regarding the shape and size of the dorsal fin, *S. plumbea* has a smaller dorsal fin, slightly falcate, less triangular in shape and sits atop a prominent and well-developed dorsal hump. While *S. chinensis* has a short dorsal fin, triangular in shape, slightly recurved and has a wide base without a basal hump, and *S. teuszii* has a similar dorsal fin shape and basal hump to *S. plumbea*, but the hump tends to be more pronounced and the fin more triangular in shape with a rounded tip. In the case of the pigmentation in these animals, it varies greatly according to geographic location: *S. plumbea* are usually dark gray with lighter ventral surface shading to off-white, with light spotting sometimes present; *S. teuszii* are similar to *S. plumbea*; *S. sahulensis* are pale gray in color with flanks shading to off-white and spotted toward the ventral surface; *S. chinensis*, specially from the southern china are pure white, often with dark spots on the body and a pinkish tinge resulting from the blood flushing during periods of high activity (Jefferson and Curry, 2015; 2016).

One of the most perplexing issues in the taxonomy of this genus has been the status of humpback dolphins inhabiting the Bay of Bengal (Eastern India, Bangladesh and Myanmar). This region is an area of overlap between *S. plumbea* and *S. chinensis*, and this population shows characteristics that are similar to both these species. The absence of a dorsal hump, the shape of the dorsal fin and extensive spotting on the body with large unpigmented areas on the sides and back, are similar to *S. chinensis*, while the darker color which is characteristic of *S. plumbea* (Muralidharan, 2013; Smith et al., 2015). However, a recent study using the mitochondrial DNA suggests that they do not group with neither *S. chinensis* nor *S. plumbea*. They are actually a highly-differentiated population that is more closely related phylogenetically to *S. sahulensis* (Amaral et al., 2017). This population possibly ranging from Bangladesh, Eastern India and Sri Lanka, and with an estimated abundance of 636 individuals (Smith et al., 2015), occurs in larger group sizes than those recorded elsewhere. With an median group size estimated of 81 to 205 individuals, it is still unknown if the social organization of this population has strong social bonds or if it is characterized by a fission/fusion society, similar to other populations in the genus *Sousa* (Amaral et al., 2017; Jefferson and Curry, 2016).

With a possible reformulation of the taxonomy at hand, the evolutionary processes for each of the species and populations are a matter of current discussion. A hypothesis has been pointed out that the genus *Sousa* has a long evolutionary history and evolved early in the delphinid evolution, being at the base of the tree around the Pliocene (McGowen, 2011). It has been indicated that the origin of the genus *Sousa* has started in eastern Australia and radiated northwards and westwards in a complex fashion over the last 8.02 Million years (Lin et al., 2010). Given this possibility, it came not as a surprise that the divergence between the Australian species and *S. chinensis* appears to occur along the Wallace Line. Wallace Line has long been known to be an important biogeographic boundary for many plants and animals. The line has been thought to be primarily a factor in evolution of terrestrial organisms, largely due to the long distance that separated Asia and Australia (Jefferson and Rosenbaum, 2014). However, some studies have been shown that this boundary can also be applied to marine organisms, and cetaceans are no exception. They have found strong evidence that dolphins of the genus *Orcaella* had split into separate species on either side of a distributional gap along the Wallace Line (Beasley et al., 2005). Like humpback dolphins, *Orcaella spp.* are coastal, shallow-water animals and it appears likely to us that for both these genera, speciation along Wallace's Line has less to do with the large distances separating these land masses in the geologic past and more to do with the relatively deep water that has long separated Southeast Asia from Australia/New Guinea (Beasley et al., 2005; Lin et al., 2010).

Nevertheless, a lot of work still needs to be done, not only to clarify exactly how many species actually exist within the genus, but also to understand the mechanisms behind their distribution, if there are other oceanographic barriers and what is the genetic basis behind these influences (Jefferson and Curry, 2015).

Importance for conservation

It is important to clarify levels of divergence and structure of these isolated populations in this genus, not only to understand the processes and the loci/genes that are involved in their isolation and speciation, but also for their present conservation status (Table 1.1). Although dolphins in general, in many human societies are thought of as 'charismatic megafauna', and therefore enjoy popular status among the general public, the unfortunate reality is that, many marine mammal populations share histories of dramatic decline due to hunting and other human impacts, and these species are no exception. Since humpback dolphins live in nearshore habitat, generally near freshwater input in developing nations heavily influenced by human activities, this makes them extremely vulnerable to fatal entanglements in fishing gear, impacts of vessel traffic and the increasing degradation of their habitat (Amaral et al., 2017).

Because so little is known about humpback populations in some areas and research work has been scant, some populations of humpback dolphins may have already been extirpated, without us even being aware of it. These vulnerable populations could benefit greatly from an improved understanding of their genetic diversity and evolution, especially in ways that can inform predictions of adaptive capacity to anthropogenic pressures and expand the toolkit for conservation policy (Cammen et al., 2016).

Table 1.1 - Summary of the differences between the species of the genus *Sousa*, with the currently conservation status for each one of them. Adapted from Jefferson and Curry, 2015.

Characteristic	***S. teuszii***	***S. plumbea***	***S. chinensis***	***S. sahulensis***
Ocean	Eastern Atlantic	Western Indian	Eastern Indian and Western Pacific	Western Pacific
Range	West Sahara to Angola	South Africa to Myanmar	East India to China and SE Asia	Southern Australia to New Guinea
IUCN Red List status				
Current	Critically Endangered	Endangered	Vulnerable	Vulnerable
External Morphology				
Dorsal hump	Prominent	Prominent	No dorsal	No dorsal
Coloration	Uniform grey with lighter belly	Uniform brownish grey with lighter belly	Mostly white as adults	Dark grey back and lighter belly, curved dorsal cape
Sexual dimorphism	Dimorphic	Dimorphic	Little or no dimorphism	Slight dimorphism

Aims

To help resolve the patterns of differentiation within genus *Sousa* and better understand the evolutionary processes behind their diversity, our study conducted a population genomic analysis of humpback dolphins. Our aim was to observe the patterns of genome-wide genetic variation and differentiation,

building up on previous population genetic studies, and to conduct a first approach over the genetic basis of speciation within this genus. In particular, we addressed the following objectives:

i) uncover the population structure and demographic history, by assessing the levels of nuclear genomic divergence between the different populations and species within the genus *Sousa*: we expect to find several genetically differentiated populations unconnected by gene flow;
ii) identify signatures of selection in the genome and candidate genes or loci with a putative role in the establishment of reproductive isolation and local adaptation: we expected to find genes linked to morphological and pigmentation traits to be under positive selection in some populations.

The results of our study should provide an important insight into the processes underlying the evolution of diversity within genus and inform on the establishment of effective conservation programs, such as the implementation of Marine Protected Areas in undeveloped areas of the Indo-Pacific.

2. Materials and Methods

Sample collection and sequencing

Our total data set consisted of 36 samples obtained from stranded or biopsied humpback dolphins, which were selected from a set of samples already used in previous studies (Amaral et al., 2017; Mendez et al., 2013a). As shown in Figure 2.1, our data set represents the entire range of the genus *Sousa*, contains samples from West Africa (WEA, n=1), Southeast Africa (SEA - South Africa, Mozambique and Tanzania, n=11), Oman (OM, n=8), Bangladesh (BAN, n=10), China (CHI - Hong Kong and Taiwan, n=3), Thailand (THA, n=1) and Australia (AUS, n=2).

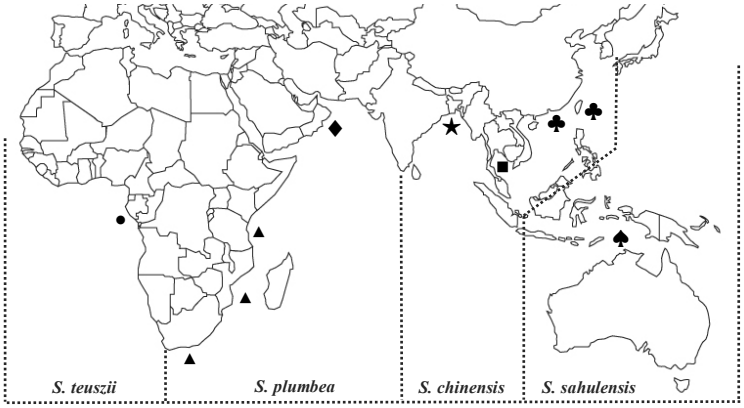

Figure 2.1 – Representation of the samples covering the entire range of the genus *Sousa*. Different symbols correspond to different populations within each species:
● – West Africa; ▲ – Southeast Africa;
♦ - Oman; ★ – Bangladesh; ■ – Thailand;
♣ – China; ♠ - Australia.

The genomic DNA from tissues samples already preserved in ethanol (96% v/v) or in sodium chloride-saturated 20% dimethyl sulphoxide (DMSO) solution, was extracted using QIAamp Tissue Kit (QIAGEN, Valencia, CA, USA) and its concentration measured using a Qubit Florometric Quantitation (ThermoFisher). The samples were then shipped to the Cornell University Institute of Biotechnology's Genomic Diversity Facility[1] where the GBS data was generated using a genotype-by-sequencing protocol (Elshire et al., 2011). To optimize the GBS results, sequencing libraries were constructed using the restriction enzyme PstI (CTGCAG), that has shown the best results in mammals (De Donato et al., 2013). Unique oligonucleotide barcodes were added to each sample for multiplexed sequencing on an Illumina HiSeq 2000 (Illumina, San Diego, CA, USA), and the template-controls were included with the batch of samples. The reads were assembled to the orca genome as a reference (*O. orca*, Oorc_1.1, 200.0x coverage, Foote et al., 2015a; Morin et al., 2010a) using bwa v0.7.8-r455 (Li and Durbin, 2009). Demultiplexing, initial quality control, assembly, and SNP discovery were completed in the TASSEL pipeline v3.0.174 (Glaubitz et al., 2014) at Cornell University.

Data processing

After the SNP calling obtained with the TASSEL pipeline, template-controls were excluded. We then applied additional filters to further reduce false positives for the following analyses. Firstly, limits for the depth of coverage were calculated and applied for each individual in RStudio v1.0.136 (RStudio Team (2016)) using a custom script. The calculation corresponded to 1/3 of the mean-depth for the minimum limit and the double of the mean-depth for the maximum limit. This calculation was applied because it considers the own average of each individual, allowing the use of more sites in the subsequent analyses. Secondly, to minimize the genotyping error that comes from heterozygosity excess, we performed a Hardy-Weinberg Equilibrium test using --hardy option in VCFtools v0.1.15 (Danecek et al., 2011) and the sites with P-values significant at the 0.01 level were excluded. Due to higher levels of

[1] http://www.biotech.cornell.edu/brc/genomic-diversity-facility

missing data in some individuals (Table 7.1), a filter of 50% missing data was conducted from VCFtools with the --max-missing option, to reduce the bias caused by inexistent data in the remaning analyses. Other filters, such as bi-allelic sites and Minimum Allele Frequency (hereafter MAF), were also applied, being conducted in VCFtools using --min-alleles, --max-alleles and --maf options, respectively.

Relatively to the MAF filter, it is already known that the MAF of the alleles tested affects the detection of a genetic effect in a given study (Glaubitz et al., 2014; Nielsen et al., 2012; Tabangin et al., 2009; Whitlock and Lotterhos, 2015). To minimize the rare alleles that are more likely to be false-positive and create bias on the results, we set three different settings for the MAF filter to produce three main data sets with the highest SNP count. The other two secondary data sets were also created for specific programs due to some technical constraints, the fourth data set for the program BEAST, while the fifth data set was used in FASTSIMCOAL2 program (Table 2.1). After this, each data set was converted to various formats using PGDSpider2 v2.1.1.3 (Lischer and Excoffier, 2012) for downstream analysis.

Table 2.1 - Summary of the five different SNP data sets that were generated for each different step. Using different values of MAF we obtain different data sets with different number of SNPs to analyze. Min and Max SNP correspond to the minimum and maximum number of loci observed within an individual, for each data set. The missing percentage corresponds to the missing data removed for each individual and the Ind. is the total individuals that data set has.

Data set	MAF (%)	Ind.	Missing (%)	Total SNPs	Min SNP	Max SNP	Singletons
1	1	32	50	25 154	5 996	22 404	12 509
2	2	32	50	19 462	4 322	16 920	6 817
3	5	32	50	11 345	2 377	10 012	491
4	2	18	50	21 103	14 759	18 937	9 284
5	5	8	100	7 090	-	7 090	0

Detection of population structure

To explore population structure in the genus *Sousa*, we first used a discriminant analysis of principal components (DAPC) to identify genetic clusters. DAPC is a multivariate approach that transforms individuals genotypes using principal components analysis (PCA) prior to a discriminant analysis (DA) (Jombart et al., 2010). This maximizes the differentiation between groups while minimizing variation within groups and was conducted using the *dapc* function in the *Adegenet* package v2.1.1 (Jombart, 2008) of the RSTUDIO framework. Since DAPC requires group assignment *a priori*, we employed a K-means clustering algorithm implemented in *Adegenet* to identify the optimal number of clusters from K = 1 to K = 10. Different clustering solutions were then compared using Bayesian Information Criterion (BIC), and to avoid over-fitting of discriminant functions, we used Alpha-score optimization to evaluate the optimal number of principal components (PCs) to retain in the analysis, as in (Jombart et al., 2010). The optimum number of PCs is indicated with the highest mean alpha across all simulations.

Second, we estimated individual genetic ancestry using sNMF (Frichot et al., 2014) through the *snmf* function in the *LEA* package v1.6.0 (Frichot and François, 2015) of the RSTUDIO framework, and the program STRUCTURE v2.3.2 (Pritchard et al., 2000). Both programs compute proportions called ancestry coefficients that represent the proportions of an individual genome that originate from multiple ancestral gene pools (Frichot et al., 2014; Pritchard et al., 2000). However, sNMF has been proven to be a faster algorithm, have comparable results to those obtained from STRUCTURE, and avoids Hardy-Weinberg equilibrium assumptions (Frichot et al., 2014) which is one of the main assumptions in STRUCTURE. Even though the advantage of sNMF has been proven in other studies, here in this work we decided to use both programs for comparative purposes.

The ancestry coefficients were estimated from a specified number of ancestral populations (K). For sNMF, the ancestry coefficient was calculated for K = 1 to K=10 using 100 replicates for each K. The preferred number of K was chosen using a cross-entropy criterion based on the prediction of masked genotypes to evaluate the error of ancestry estimation. For STRUCTURE, a correlated allele frequency model with no admixture was used (Hubisz et al., 2009). We conducted 20 runs for each K value with a burn-in of 10 000 repetitions for each value of K (Puechmaille, 2016). To determine the best value of K we employed two approaches. We used an iterative approach based on the ΔK statistic (Evanno et al., 2005) and also used the *ln(Pr(X|K))* values in order to identify the K for which *Pr(K=k)* is highest, as described in (Pritchard et al., 2000). Both approaches were calculated using CLUMPAK (Kopelman et al., 2015) and STRUCTURE HARVESTER v0.6.94 (Earl and vonHoldt, 2012).

Finally, to measure the genetic differentiation between populations, we used the *snpgdsFst* function in the SNPRelate package (Zheng et al., 2012) of the RStudio framework. The estimator of Wright's (1951) F_{ST} (henceforth F_{ST}) was calculated following the approach of Weir & Cockerham's (1984). All of these different analyses were calculated for the three main data sets and the BEAST data set, to observe the effect of MAF over population structure.

Phylogenetic relationship, Divergence time estimation and Demographic history

The phylogenetic relationships within populations were inferred with Bayesian and Maximum-likelihood methods. The Maximum-likelihood method was implemented using RAxML v8.2.11 (Stamatakis, 2014), where we inferred 1000 inferences using the ASC_GTRCAT model with no rate heterogeneity modelled (-V). The branch support was estimated using bootstrap by a majority-rule criteria as implemented in RAxML (Pattengale et al., 2010) and visualized simultaneously in a single consensus tree (Holland et al., 2005) in Figtree v1.4.3 (Rambaut et al., 2012). The consensus tree was set at 0.1, which means that bipartitions that appeared in at least 200 of the 2000 bootstrap trees participated in network construction. RAxML was ran with all the three main data sets to observed the congruency between them in the phylogenetic relationship.

For the Bayesian method we used the SNAPP package v1.3.0 (Bryant et al., 2012) that estimates species trees under a Bayesian multispecies coalescent framework through BEAST v2.4.7 (Bouckaert et al., 2014). For a priori species assignments, we used the previously identified groups from PCA and sNMF and the following parameter settings: mutation rate *u* and *v* set at 4.86 and 0.56 respectively, keeping the remaining parameters and their defaults, and a single run of the MCMC chain with 10,000,000 generations sampling every 1,000 steps was done. To assess if the posterior distribution was adequately sampled we used TRACER v1.7 (Rambaut et al., 2018) and accepted only if the effective sample size were larger than 200 for every parameter. We used DENSITREE v2.2.1 (Bouckaert and Heled, 2014) to visualize the distribution of trees, and the maximum-clade-credibility tree was generated using SNAPP Tree Set Analyser v2.4.7 (Bryant et al., 2012) with a burn-in of 10% of the trees. To obtain the divergence time estimations between the different nodes we used the substitution rate found in common bottlenose dolphin of 0.84×10^{-9} substitutions per site per year (Zhou et al., 2013). Since BEAST runs with no missing data for each population, the fourth data set was used to maximize the number of SNPs used for each population to make the best estimations possible.

The demography history was assessed by using FASTSIMCOAL2 v2.6 (Excoffier and Foll, 2011; Excoffier et al., 2013) to estimate the effective number of migrants exchanged between different populations, the effective population size for each population and also to compare the divergence time estimation for each node with the results from SNAPP. Since FASTSIMCOAL2 does not accept missing data per individual, to maximize the number of SNPs to analyze, we reduced the number of individuals to the ones with the minimum missing data, giving origin to the fifth data set (Table 2.1). Due to

computational constraints, we only analyzed three populations – Bangladesh, Oman and Australia – and three scenarios where only the order of populations was changed. The folded site frequency spectrum (SFS) was obtained through Arlequin v3.5 (Excoffier and Lischer, 2010). The mutation rate we used was 0.84 x 10 -9 substitutions per site per year (Zhou et al., 2013), and for each scenario 10 runs were carried out with the following settings: *-d -n200,000 -M -L40 -q -0*.

Model-based and Model-free selection test

It is of great interest to understand any process of selection that could have influenced the divergence of *Sousa* species. In this study, three different selection tests were performed: a Bayesian approach implemented in BAYESCAN v2.1 (Foll and Gaggiotti, 2008); a maximum likelihood test implemented in FDIST (Beaumont and Nichols, 1996); and a nonmodel-based method implemented in PCAdapt v3.0.4 (Luu et al., 2017). For the following analysis we use only the outlier SNPs that have shown in all three tests, to minimize the false positive rate that both BAYESCAN and FDIST have been shown to suffer as a result of violations of these independence assumptions (Whitlock and Lotterhos, 2015).

BAYESCAN tests whether subpopulation-specific allele frequencies, measured by an FST coefficient, are significantly different from the allele frequency within the common gene pool, and assigns a posterior probability (alpha) to a model in which selection explains a difference in allele frequencies better than a null model (Foll and Gaggiotti, 2008). A positive alpha indicates population-specific directional selection, while a negative alpha suggests balancing or purifying selection. BAYESCAN may also suffer from elevated false-positive rates under isolation by distance and range expansion, with balancing or purifying selection being especially prone to such issues (Lotterhos and Whitlock, 2014). To minimize such issues, we focused on directional selection only, and additionally we used prior odds of 10 000. Higher prior odds have been documented to help reduce the false-positive rate at the expense of identifying true loci under selection. A false discovery rate (FDR) of 0.05 was also used, keeping in mind that although this reduces the number of false positives, true signals of selection may be missed (Foll and Gaggiotti, 2008).

The FDIST test simulate the null distribution of F_{ST} for the sample sizes observed in the data, through the assumption of the island-model. Calculations were done in the program Arlequin v3.5.2.2 (Excoffier and Lischer, 2010) where the coalescent simulations were used to get a null distribution and the confidence intervals around the observed values, to see if the observed F_{ST} can be considered outliers conditioned on the global observed F_{ST} value.

The PCAdapt was used to assess structure-based selection. PCAdapt infers SNPs that are related to population structure and are candidates to local adaptation (Luu et al., 2017) based in Mahalanobis distance (1936). The corresponding program is an R package and is applied by the *pcadapt* function, using an FDR threshold of 5% with the *qvalue* package v2.6.0 (Storey et al., 2015) of the RStudio framework.

Gene identification and gene ontology enrichment analysis

The next step is to understand the process of selection that could have influenced their divergence. For that, we investigate the SNPs that showed selection using gene annotations from NCBI and Ensembl. Using the orca genome (*O. orca*, Oorc_1.1, 200.0x coverage, Foote et al., 2015; Morin et al., 2010) and through BLASTN v2.8.0+ (Eric et al., 2014) we obtained the gene information for each SNP and queried the following genomes for homologous genes: Dolphin (*T. truncatus*, NIST Tur_tru v1, 114.5x coverage), Baiji Dolphin (*L. vexillifer*, Lipotes_vexillifer_v1, 115x coverage, Zhou et al., 2013)), cow

(*Bos taurus*, Bos_taurus_UMD_3.1.1, 9x coverage), Horse, (*Equus caballus*, EquCab3.0, 88.0x coverage), Dog, (*Canis lupus*, CanFam3.1, 7x coverage), Human (*Homo sapiens*, GRCh38.p12), mouse (*Mus musculus*, GRCm38.p6). After the identification of the genes, the same genes were tested for significant enrichment of Gene Ontology (GO) categories using functional annotation clustering tool DAVID v6.8 (Huang et al., 2009a, 2009b) and PANTHER v13.1 (Mi et al., 2017). In both tools the gene list was used to search against a background of human orthologues, and to examined the significant categories with a p-value inferior to 0.05, a correction for multiple testing the Fisher's Exact with FDR was used.

3. Results

We generated GBS data of 36 individuals in which 4 individuals were excluded due to higher levels of missing data (superior to 90% - CHI12,14; WEA01) and a likely mislabeling (CHI13), producing a final data set of 32 individuals (Table 7.1) used for the downstream analysis. With 55615 SNPs obtained using the TASSEL pipeline, the final number of SNPs after filtering ranged from 7090 to 25154 depending on the percentage of MAF, missing data discarded and of the number of individuals used for each data set. A summary of the sample's sizes, types of filters, minimum and maximum SNPs per individual and singletons for each data set are found in Table 2.1.

Population structure and differentiation

As previously mentioned, the observation of a genetic effect in population genomics is very affected by the use of filters. To understand the effect of filters and to have the most plausible result out of our samples, we ran all the population structure analysis for all the different data sets described in Table 2.1. As observed in Table 2.1, the use of MAF and missing data filters reduced enormously the number of SNPs to analyze. However, for all the data sets the same population structure was observed.

The major pattern observed between all analyses was the separation between the three major clusters, segregating the three main species, *S. sahulensis*, *S. plumbea* and *S. chinensis* (Figure 3.1). Within *S. chinensis* we can observe a clearer segregation between the population of Bangladesh and Thailand samples, supporting previous studies that used the mitochondrial DNA and that showed the population of Bangladesh as highly-differentiated from other populations, with an apparent absence of gene flow between them. Although, both STRUCTURE and sNMF to showed the only individual from Thailand as a distinct cluster but rather as an individual with a mix ancestry due to the lack of samples from other populations of *S. chinensis*. This lack of material, not only affect the distinction of Thailand as another cluster in our analysis, but made it impossible to declare the Bangladesh population as a possible new species, as we will discuss below.

Some level of population structure is also observed within *S. plumbea*. However, the segregation between the African coast and the Arabian Sea population, is only detected in DAPC and sNMF analysis (Figure 3.1). This difference in detecting a slight segregation within a cluster between different programs, it is likely due to the differences in assumptions and mathematical computations. In case of sNMF the Hardy-Weinberg assumption, one of the limitations from the program STRUCTURE, is avoided and for the DAPC the variation within clusters is minimized, achieving the best discrimination of individuals in pre-defined clusters, which PCA does not takes this into account (Jombart et al., 2010; Kalinowski, 2011). These two programs, DAPC and sNMF when compared with PCA and STRUCTURE respectively, has been described that are better at detecting structured populations from fragmented data and in this study, they helped detect and confirm this slight segregation (Frichot et al., 2014; Jombart et al., 2010). Despite the fact that both statistics used to obtain the best K gave different results in the analysis of the program STRUCTURE (Evanno: K=3; Pritchard: K=4; Figure 7.2), the Evanno statistic has been described and used as the more reliable mathematical model for this program (Evanno et al., 2005). For that reason, we did not consider the K=4 as the optimal K given by the program STRUCTURE. Moreover, the percentage of variance explained and the number of optimal clusters obtained, is different between DAPC and PCA. DAPC has 67.3% of the variance explained within 5 clusters with 5 PCs used, while PCA has 55% of the variance explained with only 2 PCs within 4 clusters. With this difference between explained variances, the slight segregation within *S. plumbea* is explained by only 12.3% of the total variance from DAPC.

The pattern formed by the five clusters with the DAPC is also supported by the F_{ST} analysis (Table 3.1). Higher values of FST (between 0.58 to 0.77) are observed for all comparisons between populations. The lowest value is observed within *S. plumbea* which is also high for a dolphin species.

Table 3.1 - FST analysis using Weir and Cockerham approach, showing differentiated populations with higher values for dolphin species.

	Bangladesh	Thailand	African Coast	Arabian Sea	Australia
Bangladesh	-	0.5807863	0.7051381	0.6698441	0.7237160
Thailand	-	-	0.7018733	0.6393775	0.7565834
African Coast	-	-	-	0.3031105	0.7687165
Arabian Sea	-	-	-	-	0.722826
Australia	-	-	-	-	-

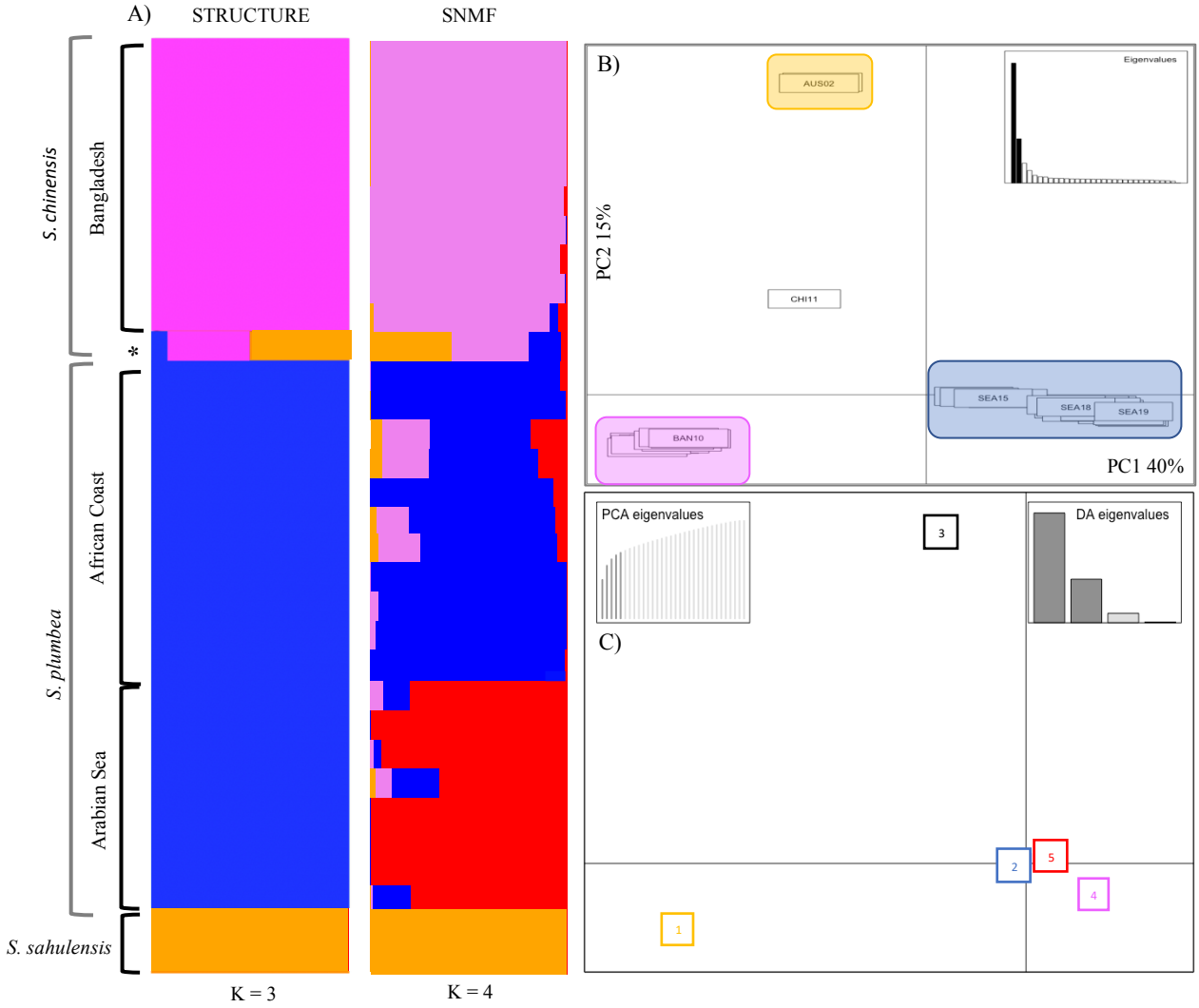

Figure 3.1 - Results from the population structure analysis of the genus *Sousa*. A) STRUCTURE and SNMF showing the clustering of different populations with little gene flow between them. B) PCA result segregating the four major clusters in this genus with 55% of the variance explained with two PC's. C) DAPC results showing five optimal clusters with 5 PCs and 4 DA eigenvalues used. Bangladesh – Pink; African Coast – Blue; Arabian Sea – Red; Australia – Yellow. Relatively to Thailand, it is represented differently for each figure: in A) is marked as *, in B) is uncoloured and C) is black.

Phylogenetic relationships and Demographic history

The Maximum Likelihood (ML) method applied through RAxML, showed the exact same pattern as the population structure analysis with the five clusters identified, supporting previous analysis (Figure 3.2). As evidenced by the consensus tree with a bootstrap highly supported in all the branches, the population of Bangladesh and Thailand are more closely related to *S. sahulensis*, with *S. plumbea* more distant from other populations studied.

To have a glimpse of the demographic history within the genus *Sousa*, we ran the SNAPP package from BEAST and FASTSIMCOAL2 with the fourth and fifth data set, respectively. Unfortunately, due to some technical adversities, both programs were unable to give statistically credible results. For SNAPP, the data set failed to be properly converted into the required nexus format creating a bias result. Besides that, the use of SNPs in this kind of methods were originally developed for DNA sequence analysis, are known to produce large posterior probability errors. As for FASTSIMCOAL2, the maximum SFS resulted in only 7000 SNPs, and considering that most of the parameters required for the models, such as times of divergence and effective population sizes were unknown, it became difficult to reach an acceptable simulation for the species in study. Additional information would likely help overcome these adversities.

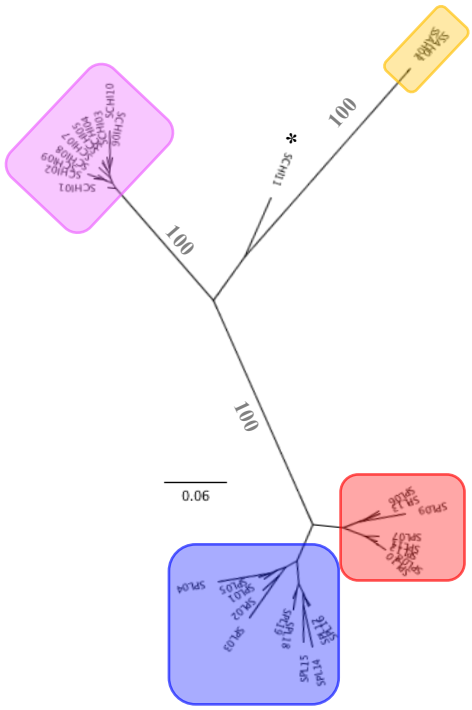

Figure 3.2 - Maximum Likelihood consensus tree obtained from RAxML with 100% of bootstrap on the longest branches. The different clusters are represented with different colours: The *S. chinensis* is separated in two clusters, the population from Bangladesh as Pink and the individual from Thailand is marked with *; *S. plumbea* separated in two clusters, the African Coast as Blue, and the Arabian Sea as Red; and the *S. sahulensis* from Australia as yellow.

Candidate genes

Within this pattern of differentiation between populations, we found evidence for genic divergence with putative functional relevance. Unfortunately, the FDIST test was not able to give credible F_{ST} observations. With heterozygosity reaching high values and with a wide distribution of F_{ST} values simulated (Figure 7.4), the assumptions of the island model used in this test most probably were violated in our data set. Without another model-based selection test to compare, to minimize the false positive rate from BAYESCAN, we ran the three main data sets and selected the SNPs that were in all of the three runs. The SNPs that appeared to be under disruptive selection according to BAYESCAN method are provided in Table 7.2. It shows the empirical outlier detection approach yielding 16 SNPs within an FDR threshold of 0.05. With *qvalues* between 0.05 to 0.01 and an alpha superior to 1, all of these 16 SNPs were consistent in the three main data sets. Relatively to PCADAPT analysis, it gave similar results.

Getting the common SNPs from the different data sets permitted us to minimized the false positive rate. The 16 SNPs also showed a correlation with the PC1 from the PCADAPT analysis, which may be associated with the segregation between *S. plumbea* to *S. chinensis* and *S. sahulensis*. The detection of this correlation with PCADAPT may be related with the higher number of individuals from *S. plumbea* and the population of Bangladesh when compared to *S. sahulensis*. From these 16 SNPs, the

corresponding genes are described in Table 7.3. In total, we found 24 genes annotations, with some of the SNPs appearing in a place of the genome that has two transcripts. Unfortunately, with only 24 genes, the functional annotation clustering from DAVID and PANTHER did not find significant functional enrichment. However, looking at the functional annotations and the tissues where each gene is expressed in humans (as described in other studies), we could observe two major roles: Brain development and Reproductive system (Table 3.2). In the following chapter we discuss some of the most interesting candidate genes and provide detailed gene descriptions for some of the genes.

Table 3.2 - A list of genes with alpha > 1 in *Sousa* populations that showed evidence of disruptive selection under development and reproductive tissues. All of these genes were described to be highly expressed in humans. The genes names and the corresponded bibliography is described in Table 8.4.

Gene	Product	Tissue*	Important role
NGEF	Nucleotide exchange factor	Brain	Brain development
FAT3	Cadherin protein	Brain	Brain development; Retinal development
TMPRSS5	Serine protease	Brain	Neuronal plasticity; Modulation of synaptic function
DRD2	Dopamine receptor	Brain	Control motor and emotional behavior
GRM7	Glutamate receptor	Brain	Brain development; Modulation of the release of glutamate and gamma-aminobutyric acid
CNTNAP5	Neurexin protein	Brain	Brain development; Implicated in cell adhesion and intercellular communication
ANO10	Transmembrane protein	Brain	Regulation of neuronal excitability
TNC	Extracellular glycoprotein	Brain, Blood vessels, Sensory Motor nerves	Brain development; Angiogenesis
GALNT15	Transferase Enzyme	Small intestine, Placenta, Brain, Ovary	Acts in O-glycosylation
NEU2	Glycohydrolytic Enzyme	Placenta, Testis and Ovary	Cell growth; Genital differentiation and Development
PAPPA	Metalloproteinase	Placenta	Modulating trophoblast invasion; Insulin-like growth factor availability; Glucose transport
CCT6B	Chaperone protein	Testis	Cytoskeletal organization; Nuclear compactation
D2HGDH	Mitochondrial Enzyme	Colon, Liver, Kidney and Brain	Epigenetic plasticity; Malignant behavior; Longevity and stem cell maintenance
GLRX3	Oxireductase enzyme	Ubiquitous	Maintaining low levels of ROS; Early embryonic growth; Involved in pregnancy-dependent mammary gland development and secretory activation
SLC36A4	Amino acid transporter	Ubiquitous	Regulation of growth and proliferation
TSEN15	Endonuclease	Ubiquitous	Brain development
MSRA	Oxireductase Enzyme		Oxidative stress resistance; Repair and regulation of protein function

* Highly expressed tissues in humans

4. Discussion

In this study we analyzed humpback dolphins occurring in the Indo-Pacific Ocean, with the aim to investigate the population structure using genome-wide markers and do a first approach on the genetic basis of population divergence in the genus. Despite some methodological issues and small sample sizes, this study supports previous studies that found differences among regions and morphotypes of humpback dolphins (Jefferson and Rosenbaum, 2014).

Highly structured genus

All studies that have analyzed the population structure and taxonomy of the genus *Sousa*, have shown that this genus has highly structured populations with none or very little gene flow between them, and with an apparent importance for geographic adaptation in genetic and morphological features. Our study supports previous findings, organizing the genus into five clusters. First, the three species already described (*S. plumbea*, *S. chinensis* and *S. sahulensis*) were clearly separated, then Bangladesh population that was highly-differentiated from the other species, and finally, the *S. plumbea* populations from the African coast and Arabian Sea also appeared segregated.

Relatively to the Bangladesh population, our results show that this population has a closer phylogenetic relationship with *S. sahulensis*, even though it shows similar characteristics with two other species, *S. plumbea* and *S. chinensis*, such as extensive spotting on the body, low prominent dorsal hump and keel at the posterior of the dorsal fin. Although, both nuclear and mitochondrial markers (Amaral et al., 2017) show the population of Bangladesh to be highly differentiated in comparison to other putative *Sousa*, the small sample size around the distribution range of *S. chinensis*, especially from Hong Kong, Taiwan and Thailand populations, makes it difficult to identify the genetic divergence of this population as a new species. Since this population is located in a region of sympatry between *S. plumbea* and *S. chinensis* and shows morphological characters between the two species, it was proposed that this population may be a case of hybridization between the two types in this area (Jefferson and Rosenbaum, 2014; Mendez et al., 2013b). However, our results do not support this hypothesis of a hybrid origin because this population, as in the mitochondrial DNA, does not group with either one of the two species. Nonetheless, there is another possible explanation. A holotype named *S. lentiginosa* (Owen,1866) discovered in Sri Lanka and Eastern India has not been confirmed as a separate species, because it was collected within the range of two species, *S. plumbea* and *S. chinensis*. This holotype has very similar morphological characteristics to the individuals from the population of Bangladesh (Figure 4.1), and it was hypothesized that this population corresponds to this species. To resolve the taxonomy around the population of Bangladesh and the *S. lentiginosa* holotype, additional information is required. Biological samples from other populations of *S. chinensis* and from *S. lentiginosa* holotype for the genetic analysis, and the corresponded morphological characteristics, will help ascertain the morphological and genetic distinction of Bangladesh population and verify if the *S. lentiginosa* is a species or classify a new species name for this population in the Bay of Bengal.

With the possibility to declare this population as a new species, the species concepts that could best be applied in this circumstance are discussed. The BSC concept, although widely used, would not be the best suited in this situation since there is no evidence that a barrier to gene flow exists between the Bangladesh population and its neighbouring populations. The PSC concept and the genic view of speciation would be more suitable. Both concepts interpret species from a genetic point of view, in that

different clusters are considered different species if they are following their own independent evolutionary trajectory, which seems to be the case of the Bangladesh humpback dolphin population.

Figure 4.1 – Wash up of *S. Lentiginosa* holotype in Sri Lanka shore. It is still unclear if this is a distinct geographic form related with the population from Bangladesh recently identified as highly-differentiated. Image adapted from Jefferson and Rosenbaum, 2014.

Candidate Genes

It is still unknown why the *Sousa* genus is so genetically structured. Some studies have hypothesized about the clumped nature of estuarine prey driven by the complex dynamics of freshwater flow, and marine currents and tides could be factors behind the genetic structuration of the genus. However, we found evidence for genic divergence of putative functional relevance, such as brain development and over the reproductive system. To minimize the potential problem of over analyzing and 'storytelling', we keep the discussion brief, and we focus only on candidate genes that have high support as outliers from multiple tests and that are convincing candidates given what we know about the populations. We analyze these genes comparing their roles and were they are highly expressed in human tissues.

Brain Genes

From the 24 genes annotated, ten of them are highly expressed in the brain region with great impact on brain development. Some of these are the genes DRD2 and GRM7, which are both receptors for different neurotransmitters located in postsynaptic neurons. DRD2 is a dopamine receptor which is involved in pathways related to reinforcement and gratification. Dopamine is the major catecholamine neurotransmitter in mammalian brain and plays important roles in diverse neurological functions, such as voluntary movement, memory formation, reward and learning (Göllner and Fieder, 2015; Obregón et al., 2017). Due to this involvement in many different processes and systems, dopamine is also related to a variety of diseases, such as Parkinson's disease, Schizophrenia, Tourette's syndrome and Attention Deficit Hyperactivity disorder (ADHD) (Rangel-Barajas et al., 2015). All of these diseases have been described not only to affect social behavior, but also locomotion and learning capability in humans. The GRM7 has a very similar activity to DRD2 and a mutation in this gene can cause the same diseases as a mutation on a Dopamine receptor. With such similarities between these two receptors, they both show similar neurological functions, such as stress, memory, reward control, circadian activity and brain emotion circuits (Niu et al., 2015). They also are predominately expressed during the brain development including neuronal migration, differentiation, synaptogenesis and neurite outgrowth (Noroozi et al., 2016).

Besides these two neuronal receptors, there are other genes that also show highly expression in the brain, and appears to complement some neuronal functions from the previous genes. ANO10 is an ion channel with a participation in regulation of neuronal excitability and when mutated can cause cerebellar ataxia, which is a disease that cause the inability to coordinate balance (Pedemonte and Galietta, 2014). CNTNAP5, FAT3 and NGEF are genes that are abundantly expressed in fetal brain, with essential roles in the correct development of the peripheral and central nervous system, that involves cell adhesion,

axon guidance and maturation of neuromuscular junctions (Chang et al., 2018; Mitsui et al., 2002; Yu et al., 2018). TMPRSS5 is a serine protease, predominantly expressed in neurons specially ate the synapses, and it modulates the synaptic function (Mitsui, 2008). Lastly, TSEN15 has demonstrated to have an important role over brain development, even though it has an ubiquitous expression (Breuss et al., 2016a).

In general, all of the ten highly expressed genes in human brain, appear to be related with many traits that affects social behaviour in humans. Social traits, especially culture features have been pointed out to driven different social systems in cetacean species. Signals of selection have also been documented in functional genes that correspond to some cultural behaviours (Foote et al., 2016; Whitehead, 2017). However, the functionality of the genes is not related with neurological functions with some effect on social behaviour. Even though, there is no documentation in cetacean species, some have been documented in humans and other animals (vonHoldt et al., 2017; Zhang-James et al., 2018) which made us hypothesize that some social biological traits are possibly being selected in any of the populations in the genus *Sousa*, has it will be discussed below.

Genes of the Reproductive System

The second most common annotation in our set of 24 genes under selection are related with the reproductive system, especially with the tissues from placenta, testis and ovaries. However, only two genes have a specific function in the reproductive system, the CCT6B and PAPPA. The CCT6B is a chaperone protein involved in protein folding that is only expressed in the testis in humans, with an important role on the organization of cytoskeletal and nuclear compaction during spermatogenesis (Agarwal et al., 2016). While PAPPA is a metalloproteinase made by the placental trophoblast cells and endometrial stromal cells, with a function of modulating trophoblast invasion, availability of insulin-growth factor and transport of glucose in the placenta (Dunne et al., 2017). There are others detected genes that show other functions beyond the reproductive system, such the NEU2 and GLRX3. The NEU2 has high expression in placenta , ovary and testis in human, pointing out its contribution to cell growth and genital differentiation and development, however in mouse NEU2 is also involved in muscle cell and neural differentiation, which is interesting because it relates with the previous described brain genes (Koseki et al., 2012; Smutova et al., 2014). GLRX3 besides is ubiquitous expression and the important role over the maintenance of reactive oxygen species, it shows also an importance over the pregnancy dependent mammary gland development and secretory activation (Pham et al., 2015, 2016).

Even though we did not find more genes specific for the reproductive system in this genus *Sousa*, in other species of cetacean it has been documented more genes also related with the testis development and acrosome reaction of the sperm (Amaral et al., 2011; Foote et al., 2016). Because of their functional role in fertilization of these genes in this study, they are emerging as candidates for a post-zygotic barrier that may be already in motion in the speciation process in some populations of the genus *Sousa*.

Social and ecological drivers

Environmental differences influence evolutionary divergence between populations, and oceanographic conditions are no different. Bay of Bengal has extraordinary oceanographic conditions, such as shallow water, intrusion of dynamic freshwater and sediment flow from the world's largest river systems, leaf litter and other bio-productivity from the world's largest mangrove forest associated with a seasonally reversing current gryre with mesoeddies (Hussain *et al.* 1994; Cheng *et al.* 2013). All together these local conditions are unique in terms of occurrence and size, and almost certainly explain the genetic distinctiveness found in most marine organisms' populations in the Bay of Bengal.

Several marine species (Ahti et al., 2016; Bowen et al., 2016; Farhadi et al., 2017; Li et al., 2015a) that occur in the Indo-West Pacific Ocean have been found to have distinct genetic lineages in the east and the west, such as those seen in the dolphin species studied here. This phylogeographic pattern may have resulted from restricted connectivity of populations across the Sunda shelf (southeast extension of the continental shelf of Southeast Asia comprising the Malay Peninsula, Sumatra, Borneo, Java and Bali) during periods of low sea level in the glacial periods of the Pleistocene (Voris, 2000). Oceanographic variables have been shown to drive population differentiation not only in humpback dolphins along the Western Indian Ocean (Mendez et al., 2011) but also in other cetacean species, such as bottlenose dolphins (Bilgmann et al., 2007), common dolphins (Amaral et al., 2012) and franciscana dolphins (Mendez et al., 2010).

However, the analysis to detect possible candidate genes showed us a possibility that social biological traits may also have an influence over the evolutionary divergence of the genus *Sousa*. Cetacea are well known for having variations in behaviors and complex social systems, in spite the fact that these animals are hard to study, there has been an increase evidence for how social structure affects their divergence. Besides that, almost all cetacean species whose behavior has been studied show possibly, or likely, culturally acquired behavior. Culture, as an inherited system, can be defined as behavior or information shared within a community that is acquired from conspecifics through some form of social learning, i.e., learning that is influenced by observation of, or interaction with, another animal or its products (Whitehead, 2017). Social learning comes in a range of forms including imitation, emulation, teaching, and local enhancement, all of which can promote behavioral similarity between learner and model. Culture may include a wide range of behavior, including foraging methods, vocalizations, diet selection, social behavior, movement, habitat use, social structure, and play (Whitehead and Rendell, 2015). The culture needs to be quite stable and to affect fitness directly or indirectly (Hoppitt and Laland, 2017). With this idea, coevolution was developed to explain how behavior is a process of two different and interacting evolutionary processes, genetic evolution and cultural evolution.

This theory has been considered entirely from the perspective of *Homo sapiens*, although culture is clearly present in other species, such as birds and cetaceans (Creanza and Feldman, 2016; Whiten et al., 2017). A great example of this are the stable, sympatric social groups with matrilineal social systems, the killer whale (*Orcinus orca*) and the sperm whale (*Physeter macrocephalus*). In these species, females spend their lives grouped with their mothers while both are alive, forming stable social units of about 10 animals. This socio-cultural relationship sets up conditions in which gene–culture coevolution could lead to neutral or functional genes being found in different elements of the social system or to more general effects such as speciation or reductions in genetic diversity (Whitehead and Rendell, 2015).

In genus *Sousa* little is known about the behaviour of the different populations, their associations are described as similar to the fission-fusion structure and do not have larger groups sizes. Although, the group size is one characteristic that has been seen to change from population to population, for example the Bangladesh and Arabian Sea populations showing group sizes up to 200 individuals, whereas the group sizes are never larger than 10 individuals in the other populations (Jefferson and Curry, 2016). Unfortunately, with this much information is hard to put forward an explanation for the observed cluster of genes related to social characteristics. There is probably unobservable behaviours such vocalizations, foraging, imitation behavior, that might explain this clustering of genes. Interestingly, all populations appear to be regionally separated, so we hypothesized that this highly structured genus may be affected by ecological drivers and social drivers may be important for the maintenance of gene pool.

5. Final considerations

By analyzing 11 345 genome-wide SNPs, the present study used population genomic analysis to evaluate the variability and differentiation in Indo-Pacific populations of the genus *Sousa*. Our work supports previous studies where five clusters were also observed. The three main species (*S. sahulensis*, *S. plumbea* and *S. chinensis*) were clearly separated from each other with absence of gene flow between them; a slightly segregation within *S. plumbea* was also observed separating the African Coast population from the Arabian Sea and the population from Bangladesh was highly-differentiated from other species with no gene flow between them. With both mtDNA and nuclear markers supporting the high differentiation of this population, and with the apparent difference in morphology and with no signs of hybridization between the two species in sympatry in this area of occurrence (*S. chinensis* and *S. plumbea*), the taxonomy of this genus appears to need a new revision. Unfortunately, it is still difficult to declare the population from Bangladesh as a new species in this study, mostly due to the lack of samples from other populations of *S. chinensis*. For future work, to resolve the taxonomy of this genus it is important to include samples within *S. chinensis* range of distribution, especially from Hong Kong, Taiwan and Thailand, to evaluate the possible structure within this species and to confirm if indeed they belong to the already holotype of *S. lentiginosa*.

With this high level of differentiation within genus *Sousa*, a number of outlier loci displaying some degree of genetic differentiation due to divergent selection were detected, with two biological functions – Brain development and Reproductive system. Most of the outlier loci had important roles affecting social biological traits. With low gene flow we found between populations, and with the genes suggested in this study, we suggest that ecological and social drives may be involved in the structure and the maintenance of the structured genus, and it may be already in motion genes with post-zygotic barrier in some populations in the genus *Sousa*.

The clarification of the population structure within the genus *Sousa*, it is not only important for the resolution of the taxonomy of the genus, but also extremely important for the conservation of these species. These species live in nearshore habitats with freshwater input, in developing nations heavily influenced by human activities, making them extremely vulnerable to fatal entanglements in fishing gear, impacts of vessel traffic and the increasing degradation of their habitat. A new evaluation of the genus, it will permit to create new conservation policies can be implemented to help minimized the anthropogenic pressures and benefit the maintenance of the diversity within this genus. This study also serves as an example of the power of population genomics approaches to uncover evidence of selection in hard to study, non-model organisms.

6. References

Agapow, P., Bininda-Emonds, O.R.P., Crandall, K.A., Gittleman, J.L., Mace, G.M., Marshall, J.C., and Purvis, A. (2004). The Impact of Species Concept on Biodiversity Studies. Q. Rev. Biol. *79*, 161–179.

Agarwal, A., Sharma, R., Samanta, L., Durairajanayagam, D., and Sabanegh, E. (2016). Proteomic signatures of infertile men with clinical varicocele and their validation studies reveal mitochondrial dysfunction leading to infertility. Asian J. Androl. *18*, 282.

Ahti, P.A., Coleman, R.R., DiBattista, J.D., Berumen, M.L., Rocha, L.A., and Bowen, B.W. (2016). Phylogeography of Indo-Pacific reef fishes: sister wrasses Coris gaimard and C. cuvieri in the Red Sea, Indian Ocean and Pacific Ocean. J. Biogeogr. *43*, 1103–1115.

Amaral, A.R., Möller, L.M., Beheregaray, L.B., and Coelho, M.M. (2011). Evolution of 2 Reproductive Proteins, ZP3 and PKDREJ, in Cetaceans. J. Hered. *102*, 275–282.

Amaral, A.R., Beheregaray, L.B., Bilgmann, K., Boutov, D., Freitas, L., Robertson, K.M., Sequeira, M., Stockin, K.A., Coelho, M.M., and Möller, L.M. (2012). Seascape Genetics of a Globally Distributed, Highly Mobile Marine Mammal: The Short-Beaked Common Dolphin (Genus Delphinus). PLoS ONE *7*, e31482.

Amaral, A.R., Smith, B.D., Mansur, R.M., Brownell, R.L., and Rosenbaum, H.C. (2016). Oceanographic drivers of population differentiation in Indo-Pacific bottlenose (Tursiops aduncus) and humpback (Sousa spp.) dolphins of the northern Bay of Bengal. Conserv. Genet. *18*, 371–381.

Amaral, A.R., Smith, B.D., Mansur, R.M., Brownell, R.L., and Rosenbaum, H.C. (2017). Oceanographic drivers of population differentiation in Indo-Pacific bottlenose (Tursiops aduncus) and humpback (Sousa spp.) dolphins of the northern Bay of Bengal. Conserv. Genet. *18*, 371–381.

Attard, C.R.M., Beheregaray, L.B., Sandoval-Castillo, J., Jenner, K.C.S., Gill, P.C., Jenner, M.-N.M., Morrice, M.G., and Möller, L.M. (2018). From conservation genetics to conservation genomics: a genome-wide assessment of blue whales (*Balaenoptera musculus*) in Australian feeding aggregations. R. Soc. Open Sci. *5*, 170925.

Beasley, I., Robertson, K.M., and Arnold, P. (2005). Description of a new dolphin, the Australian Snubfin Dolphin Orcaella Heinsohni sp. n. (Cetacea, Delphinidae). Mar. Mammal Sci. *21*, 365–400.

Beaumont, M.A., and Nichols, R.A. (1996). Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B *263*, 1619–1626.

Bilgmann, K., Möller, L., Harcourt, R., Gibbs, S., and Beheregaray, L. (2007). Genetic differentiation in bottlenose dolphins from South Australia: association with local oceanography and coastal geography. Mar. Ecol. Prog. Ser. *341*, 265–276.

Binda, A.V., Kabbani, N., Lin, R., and Levenson, R. (2002). D2 and D3 Dopamine Receptor Cell Surface Localization Mediated by Interaction with Protein 4.1N. Mol. Pharmacol. *62*, 507–513.

Bouckaert, R., and Heled, J. (2014). DensiTree 2: Seeing Trees Through the Forest.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLOS Comput. Biol. *10*, e1003537.

Bowen, B.W., Gaither, M.R., DiBattista, J.D., Iacchei, M., Andrews, K.R., Grant, W.S., Toonen, R.J., and Briggs, J.C. (2016). Comparative phylogeography of the ocean planet. Proc. Natl. Acad. Sci. *113*, 7962–7969.

Breuss, M.W., Sultan, T., James, K.N., Rosti, R.O., Scott, E., Musaev, D., Furia, B., Reis, A., Sticht, H., Al-Owain, M., et al. (2016a). Autosomal-Recessive Mutations in the tRNA Splicing Endonuclease Subunit TSEN15 Cause Pontocerebellar Hypoplasia and Progressive Microcephaly. Am. J. Hum. Genet. *99*, 228–235.

Breuss, M.W., Sultan, T., James, K.N., Rosti, R.O., Scott, E., Musaev, D., Furia, B., Reis, A., Sticht, H., Al-Owain, M., et al. (2016b). Autosomal-Recessive Mutations in the tRNA Splicing Endonuclease Subunit TSEN15 Cause Pontocerebellar Hypoplasia and Progressive Microcephaly. Am. J. Hum. Genet. *99*, 228–235.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., and RoyChoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. Mol. Biol. Evol. *29*, 1917–1932.

Budde, B.S., Namavar, Y., Barth, P.G., Poll-The, B.T., Nürnberg, G., Becker, C., Van Ruissen, F., Weterman, M.A.J., Fluiter, K., Te Beek, E.T., et al. (2008). tRNA splicing endonuclease mutations cause pontocerebellar hypoplasia. Nat. Genet. *40*, 1113–1118.

Cammen, K.M., Andrews, K.R., Carroll, E.L., Foote, A.D., Humble, E., Khudyakov, J.I., Louis, M., McGowen, M.R., Olsen, M.T., and Van Cise, A.M. (2016). Genomic Methods Take the Plunge: Recent Advances in High-Throughput Sequencing of Marine Mammals. J. Hered. *107*, 481–495.

Chang, C.J., Chang, M.Y., Chou, S.Y., Huang, C.C., Chuang, J.Y., Hsu, T.I., Chang, H.F., Wu, Y.H., Wu, C.C., Morales, D., et al. (2018). Ephexin1 is required for Eph-Mediated limb trajectory of spinal motor axons. J. Neurosci. *38*, 2043–2056.

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. Genome Res. *20*, 393–402.

Cheng, H., Burroughs-Garcia, J., Birkness, J.E., Trinidad, J.C., and Deans, M.R. (2016). Disparate Regulatory Mechanisms Control Fat3 and P75NTR Protein Transport through a Conserved Kif5-Interaction Domain. PLOS ONE *11*, e0165519.

Cheng, L., Tachibana, K., Iwasaki, H., Kameyama, A., Zhang, Y., Kubota, T., Hiruma, T., Tachibana, K., Kudo, T., Guo, J.-M., et al. (2004). Characterization of a novel human UDP-GalNAc transferase, pp-GalNAc-T15. FEBS Lett. *566*, 17–24.

Cheng, X., Xie, S.-P., McCreary, J.P., Qi, Y., and Du, Y. (2013). Intraseasonal variability of sea surface height in the Bay of Bengal: SEA SURFACE HEIGHT ISV IN THE BoB. J. Geophys. Res. Oceans *118*, 816–830.

Coyne, J.A. (2010). Why evolution is true (New York: Penguin Books).

Creanza, N., and Feldman, M.W. (2016). Worldwide genetic and cultural change in human evolution. Curr. Opin. Genet. Dev. *41*, 85–92.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

De Donato, M., Peters, S.O., Mitchell, S.E., Hussain, T., and Imumorin, I.G. (2013). Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing. PLoS ONE *8*, e62137.

Dunne, C., Cho, K., Shan, A., Hutcheon, J., Durland, U.S., Seethram, K., and Havelock, J.C. (2017). Peak Serum Estradiol Level During Controlled Ovarian Stimulation Is not Associated with Lower Levels of Pregnancy-Associated Plasma Protein-A or Small for Gestational Age Infants: A Cohort Study. J. Obstet. Gynaecol. Can. *39*, 870–879.

Earl, D.A., and vonHoldt, B.M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. *4*, 359–361.

Eldredge, N., and Cracraft, J. (1980). Phylogenetic patterns and the evolutionary process: method and theory in comparative biology (New York: Columbia University Press).

Ellegren, H., Smeds, L., Burri, R., Olason, P.I., Backström, N., Kawakami, T., Künstner, A., Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., et al. (2012). The genomic landscape of species divergence in Ficedula flycatchers. Nature *491*, 756–760.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE *6*, e19379.

Eric, S.D., Nicholas, T.K.D.D., and Theophilus, K.A. (2014). Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). J. Bioinforma. Seq. Anal. *6*, 1–6.

Etter, P.D., Bassham, S., Hohenlohe, P.A., Johnson, E.A., and Cresko, W.A. (2012). SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing. In Molecular Methods for Evolutionary Genetics, V. Orgogozo, and M.V. Rockman, eds. (Totowa, NJ: Humana Press), pp. 157–178.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. *14*, 2611–2620.

Excoffier, L., and Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics *27*, 1332–1334.

Excoffier, L., and Lischer, H.E.L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. *10*, 564–567.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., and Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. PLOS Genet. *9*, e1003905.

Farhadi, A., Jeffs, A.G., Farahmand, H., Rejiniemon, T.S., Smith, G., and Lavery, S.D. (2017). Mechanisms of peripheral phylogeographic divergence in the indo-Pacific: lessons from the spiny lobster Panulirus homarus. BMC Evol. Biol. *17*.

Feder, J.L., Egan, S.P., and Nosil, P. (2012). The genomics of speciation-with-gene-flow. Trends Genet. *28*, 342–350.

Foll, M., and Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. Genetics *180*, 977–993.

Foote, A.D., Liu, Y., Thomas, G.W.C., Vinař, T., Alföldi, J., Deng, J., Dugan, S., van Elk, C.E., Hunter, M.E., Joshi, V., et al. (2015). Convergent evolution of the genomes of marine mammals. Nat. Genet. *47*, 272–275.

Foote, A.D., Vijay, N., Ávila-Arcos, M.C., Baird, R.W., Durban, J.W., Fumagalli, M., Gibbs, R.A., Hanson, M.B., Korneliussen, T.S., Martin, M.D., et al. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. Nat. Commun. *7*, 11693.

Frichot, E., and François, O. (2015). LEA : An R package for landscape and ecological association studies. Methods Ecol. Evol. *6*, 925–929.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics *196*, 973–983.

Gardner, M., Bertranpetit, J., and Comas, D. (2008). Worldwide genetic variation in dopamine and serotonin pathway genes: Implications for association studies. Am. J. Med. Genet. B Neuropsychiatr. Genet. *147B*, 1070–1075.

Gatesy, J., Geisler, J.H., Chang, J., Buell, C., Berta, A., Meredith, R.W., Springer, M.S., and McGowen, M.R. (2013). A phylogenetic blueprint for a modern whale. Mol. Phylogenet. Evol. *66*, 479–506.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., and Buckler, E.S. (2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. PLoS ONE *9*, e90346.

Göllner, T., and Fieder, M. (2015). Selection in the dopamine receptor 2 gene: a candidate SNP study. PeerJ *3*, e1149.

Gong, J., Tian, J., Lou, J., Wang, X., Ke, J., Li, J., Yang, Y., Gong, Y., Zhu, Y., Zou, D., et al. (2018). A polymorphic MYC response element in KBTBD11 influences colorectal cancer risk, especially in interaction with an MYC-regulated SNP rs6983267. Ann. Oncol. *29*, 632–639.

Haasl, R.J., and Payseur, B.A. (2016). Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. Mol. Ecol. *25*, 5–23.

Han, J., Jackson, D., Holm, J., Turner, K., Ashcraft, P., Wang, X., Cook, B., Arning, E., Genta, R.M., Venuprasad, K., et al. (2018). Elevated d-2-hydroxyglutarate during colitis drives progression to colorectal cancer. Proc. Natl. Acad. Sci. *115*, 1057–1062.

He, F., Wei, L., Luo, W., Liao, Z., Li, B., Zhou, X., Xiao, X., You, J., Chen, Y., Zheng, S., et al. (2016). Glutaredoxin 3 promotes nasopharyngeal carcinoma growth and metastasis via EGFR/Akt pathway and independent of ROS. Oncotarget *7*, 37000–37012.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., and Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front. Plant Sci. *5*.

Holland, B.R., Delsuc, F., Moulton, V., and Baker, A. (2005). Visualizing Conflicting Evolutionary Hypotheses in Large Collections of Trees: Using Consensus Networks to Study the Origins of Placentals and Hexapods. Syst. Biol. *54*, 66–76.

Hoppitt, W., and Laland, K.N. (2017). Social learning: an introduction to mechanisms, methods, and models.

Hu, B., and El Haj, A.J. (2013). Methionine sulfoxide reductase A as a marker for isolating subpopulations of stem and progenitor cells used in regenerative medicine. Med. Hypotheses *80*, 663–665.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the

comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Hubisz, M.J., Falush, D., Stephens, M., and Pritchard, J.K. (2009). Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Resour. *9*, 1322–1332.

Hussain Z, Acharya G (1994) Mangroves of the Sundarbans, vol 2. IUCN, Bangkok

Jefferson, T.A., and Curry, B.E. (2015). In Advances in Marine Biology: Humpback dolphins (Sousa spp.): current status and conservation part I, (Elsevier), pp. 1–16.

Jefferson, T.A., and Curry, B.E. (2016). In Advances in Marine Biology: Humpback dolphins (Sousa spp.): current status and conservation part I, (Elsevier), pp.1-16.

Jefferson, T.A., and Karczmarski, L. (2001). Sousa chinensis. Mamm. Species *655*, 1–9.

Jefferson, T.A., and Rosenbaum, H.C. (2014). Taxonomic revision of the humpback dolphins (*Sousa* spp.), and description of a new species from Australia. Mar. Mammal Sci. *30*, 1494–1541.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics *24*, 1403–1405.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. *11*, 94.

Kalinowski, S.T. (2011). The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. Heredity *106*, 625–632.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., and Mayrose, I. (2015). CLUMPAK : a program for identifying clustering modes and packaging population structure inferences across *K*. Mol. Ecol. Resour. *15*, 1179–1191.

Koseki, K., Wada, T., Hosono, M., Hata, K., Yamaguchi, K., Nitta, K., and Miyagi, T. (2012). Human cytosolic sialidase NEU2-low general tissue expression but involvement in PC-3 prostate cancer cell survival. Biochem. Biophys. Res. Commun. *428*, 142–149.

Krol, A., Henle, S.J., and Goodrich, L.V. (2016). Fat3 and Ena/VASP proteins influence the emergence of asymmetric cell morphology in the developing retina. Development *143*, 2172–2182.

Langerhans, R.B., and Riesch, R. (2013). Speciation by selection: A framework for understanding ecology's role in speciation. Curr. Zool. *59*, 31–52.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, C., Corrigan, S., Yang, L., Straube, N., Harris, M., Hofreiter, M., White, W.T., and Naylor, G.J.P. (2015a). DNA capture reveals transoceanic gene flow in endangered river sharks. Proc. Natl. Acad. Sci. *112*, 13302–13307.

Li, J., Zhao, X., Xin, Q., Shan, S., Jiang, B., Jin, Y., Yuan, H., Dai, P., Xiao, R., Zhang, Q., et al. (2015b). Whole-Exome Sequencing Identifies a Variant in TMEM132E Causing Autosomal-Recessive Nonsyndromic Hearing Loss DFNB99. Hum. Mutat. *36*, 98–105.

Lin, A.-P., Abbas, S., Kim, S.-W., Ortega, M., Bouamar, H., Escobedo, Y., Varadarajan, P., Qin, Y., Sudderth, J., Schulz, E., et al. (2015). D2HGDH regulates alpha-ketoglutarate levels and dioxygenase function by modulating IDH2. Nat. Commun. *6*, 7768.

Lin, D., Liang, Y., Zheng, D., Chen, Y., Jing, X., Lei, M., Zeng, Z., Zhou, T., Wu, X., Peng, S., et al. (2018). Novel biomolecular information in rotenone-induced cellular model of Parkinson's disease. Gene *647*, 244–260.

Lin, W., Zhou, R., Porter, L., Chen, J., and Wu, Y. (2010). Evolution of Sousa chinensis: A scenario based on mitochondrial DNA study. Mol. Phylogenet. Evol. *57*, 907–911.

Lischer, H.E.L., and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics *28*, 298–299.

Liu, Y., Lear, T., Iannone, O., Shiva, S., Corey, C., Rajbhandari, S., Jerome, J., Chen, B.B., and Mallampalli, R.K.

(2015). The Proapoptotic F-box Protein Fbxl7 Regulates Mitochondrial Function by Mediating the Ubiquitylation and Proteasomal Degradation of Survivin. J. Biol. Chem. *290*, 11843–11852.

Lotterhos, K.E., and Whitlock, M.C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Mol. Ecol. *23*, 2178–2192.

Luu, K., Bazin, E., and Blum, M.G.B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol. Ecol. Resour. *17*, 67–77.

Mallet, J. (1995). A species definition for the modern synthesis. Trends Ecol. Evol. *10*, 294–299.

Marques, D.A., Lucek, K., Haesler, M.P., Feller, A.F., Meier, J.I., Wagner, C.E., Excoffier, L., and Seehausen, O. (2017). Genomic landscape of early ecological speciation initiated by selection on nuptial colour. Mol. Ecol. *26*, 7–24.

Mavárez, J., Salazar, C.A., Bermingham, E., Salcedo, C., Jiggins, C.D., and Linares, M. (2006). Speciation by hybridization in Heliconius butterflies. Nature *441*, 868–871.

Mayr, E. (1942). Systematics and the origin of species (New York: Columbia University Press).

McGowen, M.R. (2011). Toward the resolution of an explosive radiation—A multilocus phylogeny of oceanic dolphins (Delphinidae). Mol. Phylogenet. Evol. *60*, 345–357.

Mendez, M., Rosenbaum, H.C., Wells, R.S., Stamper, A., and Bordino, P. (2010). Genetic Evidence Highlights Potential Impacts of By-Catch to Cetaceans. PLoS ONE *5*, e15550.

Mendez, M., Subramaniam, A., Collins, T., Minton, G., Baldwin, R., Berggren, P., Särnblad, A., Amir, O.A., Peddemors, V.M., Karczmarski, L., et al. (2011). Molecular ecology meets remote sensing: environmental drivers to population structure of humpback dolphins in the Western Indian Ocean. Heredity *107*, 349–361.

Mendez, M., Jefferson, T.A., Kolokotronis, S.O., Krützen, M., Parra, G.J., Collins, T., Minton, G., Baldwin, R., Berggren, P., Särnblad, A., et al. (2013a). Integrating multiple lines of evidence to better understand the evolutionary divergence of humpback dolphins along their entire distribution range: A new dolphin species in Australian waters? Mol. Ecol. *22*, 5936–5948.

Mendez, M., Jefferson, T.A., Kolokotronis, S.-O., Krützen, M., Parra, G.J., Collins, T., Minton, G., Baldwin, R., Berggren, P., Särnblad, A., et al. (2013b). Integrating multiple lines of evidence to better understand the evolutionary divergence of humpback dolphins along their entire distribution range: a new dolphin species in Australian waters? Mol. Ecol. *22*, 5936–5948.

Meyer, B.S., Matschiner, M., and Salzburger, W. (2016). Disentangling Incomplete Lineage Sorting and Introgression to Refine Species-Tree Estimates for Lake Tanganyika Cichlid Fishes. Syst. Biol. syw069.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. *45*, D183–D189.

Mi, Z., Halfter, W., Abrahamson, E.E., Klunk, W.E., Mathis, C.A., Mufson, E.J., and Ikonomovic, M.D. (2016). Tenascin-C Is Associated with Cored Amyloid-β Plaques in Alzheimer Disease and Pathology Burdened Cognitively Normal Elderly. J. Neuropathol. Exp. Neurol. *75*, 868–876.

Miglietta, M.P., Faucci, A., and Santini, F. (2011). Speciation in the Sea: Overview of the Symposium and Discussion of Future Directions. Integr. Comp. Biol. *51*, 449–455.

Mitsui, S. (2008). Mosaic serine proteases in the mammalian central nervous system. Front. Biosci. *13*, 1991.

Mitsui, K., Nakajima, D., Ohara, O., and Nakayama, M. (2002). Mammalian fat3: A Large Protein That Contains Multiple Cadherin and EGF-like Motifs. Biochem. Biophys. Res. Commun. *290*, 1260–1266.

Miura, R., Araki, A., Miyashita, C., Kobayashi, S., Kobayashi, S., Wang, S.-L., Chen, C.-H., Miyake, K., Ishizuka, M., Iwasaki, Y., et al. (2018). An epigenome-wide study of cord blood DNA methylations in relation to prenatal perfluoroalkyl substance exposure: The Hokkaido study. Environ. Int. *115*, 21–28.

Momigliano, P., Jokinen, H., Fraimout, A., Florin, A.-B., Norkko, A., and Merilä, J. (2017). Extraordinarily rapid speciation in a marine fish. Proc. Natl. Acad. Sci. *114*, 6074–6079.

Morin, P.A., Archer, F.I., Foote, A.D., Vilstrup, J., Allen, E.E., Wade, P., Durban, J., Parsons, K., Pitman, R., Li, L., et al. (2010). Complete mitochondrial genome phylogeographic analysis of killer whales (Orcinus orca) indicates multiple species. Genome Res. *20*, 908–916.

Muralidharan, R. (2013). Sightings and behavioral observations of Indo-Pacific Humpback Dolphins Sousa chinensis (Osbeck, 1765) along Chennai coast, Bay of Bengal. J. Threat. Taxa *5*, 5002–5006.

Narum, S.R., Buerkle, C.A., Davey, J.W., Miller, M.R., and Hohenlohe, P.A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. Mol. Ecol. *22*, 2841–2847.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. PLoS ONE *7*, e37558.

Niu, W., Huang, X., Yu, T., Chen, S., Li, X., Wu, X., Cao, Y., Zhang, R., Bi, Y., Yang, F., et al. (2015). Association study of GRM7 polymorphisms and schizophrenia in the Chinese Han population. Neurosci. Lett. *604*, 109–112.

Noh, M.R., Kim, K.Y., Han, S.J., Kim, J.I., Kim, H.-Y., and Park, K.M. (2017). Methionine Sulfoxide Reductase A Deficiency Exacerbates Cisplatin-Induced Nephrotoxicity via Increased Mitochondrial Damage and Renal Cell Death. Antioxid. Redox Signal. *27*, 727–741.

Noroozi, R., Taheri, M., Movafagh, A., Mirfakhraie, R., Solgi, G., Sayad, A., Mazdeh, M., and Darvish, H. (2016). Glutamate receptor, metabotropic 7 (GRM7) gene variations and susceptibility to autism: A case–control study. Autism Res. *9*, 1161–1168.

Obregón, A.M., Valladares, M., and Goldfield, G. (2017). Association of the dopamine D2 receptor rs1800497 polymorphism and eating behavior in Chilean children. Nutrition *35*, 139–145.

Orr, H.A., and Turelli, M. (2001). The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. Evol. Int. J. Org. Evol. *55*, 1085–1094.

Ozaki, K., Kuroki, T., Hayashi, S., and Nakamura, Y. (1996). Isolation of Three Testis-Specific Genes (TSA303, TSA806, TSA903) by a Differential mRNA Display Method. Genomics *36*, 316–319.

Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R.P., Moret, B.M.E., and Stamatakis, A. (2010). How Many Bootstrap Replicates Are Necessary? J. Comput. Biol. *17*, 337–354.

Pedemonte, N., and Galietta, L.J.V. (2014). Structure and Function of TMEM16 Proteins (Anoctamins). Physiol. Rev. *94*, 419–459.

Petry, C.J., Ong, K.K., Hughes, I.A., Acerini, C.L., Frystyk, J., and Dunger, D.B. (2017). Early Pregnancy-Associated Plasma Protein A Concentrations Are Associated With Third Trimester Insulin Sensitivity. J. Clin. Endocrinol. Metab. *102*, 2000–2008.

Pham, K., Pal, R., Qu, Y., Liu, X., Yu, H., Shiao, S.L., Wang, X., O'Brian Smith, E., Cui, X., Rodney, G.G., et al. (2015). Nuclear glutaredoxin 3 is critical for protection against oxidative stress-induced cell death. Free Radic. Biol. Med. *85*, 197–206.

Pham, K., Dong, J., Jiang, X., Qu, Y., Yu, H., Yang, Y., Olea, W., Marini, J.C., Chan, L., Wang, J., et al. (2016). Loss of glutaredoxin 3 impedes mammary lobuloalveolar development during pregnancy and lactation. Am. J. Physiol.-Endocrinol. Metab. *312*, E136–E149.

Pillai, S.M., and Meredith, D. (2011). SLC36A4 (hPAT4) Is a High Affinity Amino Acid Transporter When Expressed in Xenopus laevis Oocytes. J. Biol. Chem. *286*, 2455–2460.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.

Puechmaille, S.J. (2016). The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. Mol. Ecol. Resour. *16*, 608–627.

Rambaut, A. (2012). FigTree v. 1.4.0. *http://tree.bio.ed.ac.uk/software/figtree/*

Rambaut, A., Drummond, A.J., Xie, D., Baele, G., Suchard, M.A., and Susko, E. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. *67*, 901–904.

Rangel-Barajas, C., Coronel, I., and Florán, B. (2015). Dopamine Receptors and Neurodegeneration. Aging Dis. *6*, 349.

Ravinet, M., Faria, R., Butlin, R.K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M.A.F., Mehlig, B., and Westram, A.M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J. Evol. Biol. *30*, 1450–1477.

Rodrigues, N.R., Theodosiou, A.M., Nesbit, M.A., Campbell, L., Tandle, A.T., Saranath, D., and Davies, K.E.

(2000). Characterization of Ngef, a Novel Member of the Dbl Family of Genes Expressed Predominantly in the Caudate Nucleus. Genomics *65*, 53–61.

Rzechonek, A., Grzegrzolka, J., Blasiak, P., Ornat, M., Piotrowska, A., Nowak, A., and Dziegiel, P. (2018). Correlation of Expression of Tenascin C and Blood Vessel Density in Non-small Cell Lung Cancers. Anticancer Res. *38*, 1987–1991.

Schwartz, M.K., and Boness, D.J. (2017). Marine mammal subspecies in the age of genetics: Introductory remarks from the Associate Editor and Editor-in-Chief of *Marine Mammal Science*. Mar. Mammal Sci. *33*, 7–11.

Scordato, E.S.C., Symes, L.B., Mendelson, T.C., and Safran, R.J. (2014). The Role of Ecology in Speciation by Sexual Selection: A Systematic Empirical Review. J. Hered. *105*, 782–794.

Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L., Saetre, G.-P., Bank, C., Brännström, Å., et al. (2014). Genomics and the origin of species. Nat. Rev. Genet. *15*, 176–192.

Shi, L., Butt, B., Ip, F.C.F., Dai, Y., Jiang, L., Yung, W.-H., Greenberg, M.E., Fu, A.K.Y., and Ip, N.Y. (2010). Ephexin1 is required for structural maturation and neurotransmission at the neuromuscular junction. Neuron *65*, 204–216.

Simpson, G.G. (1951). The Species Concept. Evolution *5*, 285–298.

Singh, M.P., Kim, K.Y., Kwak, G.-H., Baek, S.-H., and Kim, H.-Y. (2017). Methionine sulfoxide reductase A protects against lipopolysaccharide-induced septic shock via negative regulation of the proinflammatory responses. Arch. Biochem. Biophys. *631*, 42–48.

Smith, B.D., Mansur, R., Strindberg, S., Redfern, J., and Moore, T. (2015). Population demographics, habitat selection, and a spatial and photographic analysis of bycatch risk of Indo-Pacific humpback dolphins Sousa chinensis and bottlenose dolphins Tursiops aduncus in the northern Bay of Bengal, Bangladesh.

Smutova, V., Albohy, A., Pan, X., Korchagina, E., Miyagi, T., Bovin, N., Cairo, C.W., and Pshezhetsky, A.V. (2014). Structural Basis for Substrate Specificity of Mammalian Neuraminidases. PLOS ONE *9*, e106320.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Steeman, M.E., Hebsgaard, M.B., Fordyce, R.E., Ho, S.Y.W., Rabosky, D.L., Nielsen, R., Rahbek, C., Glenner, H., Sørensen, M.V., and Willerslev, E. (2009). Radiation of Extant Cetaceans Driven by Restructuring of the Oceans. Syst. Biol. *58*, 573–585.

Storey, J. D., Bass, A. J., Dabney, A., Robinson, D. (2015). *qvalue*: Q-value estimation for false discovery rate control. R package version 2.6.0. http://github.com/jdstorey/qvalue

Tabangin, M.E., Woo, J.G., and Martin, L.J. (2009). The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc. *3*, S41.

Traut, W., Weichenhan, D., Himmelbauer, H., and Winking, H. (2006). New members of the neurexin superfamily: multiple rodent homologues of the human CASPR5 gene. Mamm. Genome *17*, 723–731.

Turelli, M., Barton, N.H., and Coyne, J.A. (2001). Theory and speciation. Trends Ecol. Evol. *16*, 330–343.

Valdivieso, P., Toigo, M., Hoppeler, H., and Flück, M. (2017). T/T homozygosity of the tenascin-C gene polymorphism rs2104772 negatively influences exercise-induced angiogenesis. PLOS ONE *12*, e0174864.

Valen, L.V. (1976). Ecological Species, Multispecies, and Oaks. Taxon *25*, 233.

vonHoldt, B.M., Shuldiner, E., Koch, I.J., Kartzinel, R.Y., Hogan, A., Brubaker, L., Wanser, S., Stahler, D., Wynne, C.D.L., Ostrander, E.A., et al. (2017). Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs. Sci. Adv. *3*, e1700398.

Voris, H.K. (2000). Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. J. Biogeogr. *27*, 1153–1167.

W. Wolf, J.B., and Ellegren, H. (2016). Making sense of genomic islands of differentiation in light of speciation. Nat. Publ. Group.

Wang, J.Y., Chu Yang, S., Hung, S.K., and Jefferson, T.A. (2007). Distribution, abundance and conservation status of the eastern Taiwan Strait population of Indo-Pacific humpback dolphins, Sousa chinensis. Mammalia *71*.

Waples, R.S., and Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods

for identifying the number of gene pools and their degree of connectivity: WHAT IS A POPULATION? Mol. Ecol. *15*, 1419–1439.

Whitehead, H. (2017). Gene–culture coevolution in whales and dolphins. Proc. Natl. Acad. Sci. *114*, 7814–7821.

Whitehead, H., and Rendell, L. (2015). The cultural lives of whales and dolphins (Chicago and London: The University of Chicago Press).

Whiten, A., Ayala, F.J., Feldman, M.W., and Laland, K.N. (2017). The extension of biology through culture. Proc. Natl. Acad. Sci. *114*, 7775–7781.

Whitlock, M.C., and Lotterhos, K.E. (2015). Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F_{ST}. Am. Nat. *186*, S24–S36.

Wolf, J.B.W., and Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. Nat. Rev. Genet. *18*, 87–100.

Wu, C.-I. (2001). The genic view of the process of speciation: Genic view of the process of speciation. J. Evol. Biol. *14*, 851–865.

Würsig, B.G., Thewissen, J.G.M., and Kovacs, K.M. (2018). Encyclopedia of marine mammals.

Yamada, K.H., Kang, H., and Malik, A.B. (2017). Antiangiogenic Therapeutic Potential of Peptides Derived from the Molecular Motor KIF13B that Transports VEGFR2 to Plasmalemma in Endothelial Cells. Am. J. Pathol. *187*, 214–224.

Yamaguchi, N., Okui, A., Yamada, T., Nakazato, H., and Mitsui, S. (2002). Spinesin/TMPRSS5, a Novel Transmembrane Serine Protease, Cloned from Human Spinal Cord. J. Biol. Chem. *277*, 6806–6812.

Yu, H., Yan, H., Wang, L., Li, J., Tan, L., Deng, W., Chen, Q., Yang, G., Zhang, F., Lu, T., et al. (2018). Five novel loci associated with antipsychotic treatment response in patients with schizophrenia: a genome-wide association study. Lancet Psychiatry *5*, 327–338.

Zhang-James, Y., Fernàndez-Castillo, N., Hess, J.L., Malki, K., Glatt, S.J., Cormand, B., and Faraone, S.V. (2018). An integrated analysis of genes and functional pathways for aggression in human and rodent models. Mol. Psychiatry.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328.

Zhou, X., Sun, F., Xu, S., Fan, G., Zhu, K., Liu, X., Chen, Y., Shi, C., Yang, Y., Huang, Z., et al. (2013). Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations. Nat. Commun. *4*, 2708.

7. Supplementary Information

Table 7.1 - Representation of all individuals missing data in percentage before any filter application. The 36 individuals are ordered by species and location, and are identified by the location and the order of position in lane of sequencing. The values of missing data vary depending on the quality of the samples for each of the individuals.

Species	Local		Ind.	Missing sites (%)
S. chinensis	Bangladesh		BAN01	3,7%
			BAN02	6,4%
			BAN03	8,8%
			BAN04	9,7%
			BAN05	4,4%
			BAN06	10,4%
			BAN07	11,8%
			BAN08	13,5%
			BAN09	18,3%
			BAN10	40,9%
	China	Thailand	CHI11	5,1%
		Taiwan	CHI12	99,4%
			CHI13	72,6%
		Hong Kong	CHI14	99,0%
S. plumbea	Southeast Africa	Tanzania	SEA01	27,2%
			SEA02	41,8%
			SEA03	79,1%
			SEA04	80,9%
			SEA05	25,2%
S. plumbea	Southeast Africa	South Africa	SEA14	66,3%
			SEA15	72,4%
			SEA16	20,6%
			SEA17	35,7%
		Mozambique	SEA18	31,7%
			SEA19	9,1%
	Oman		OM06	22,4%
			OM07	4,9%
			OM08	15,6%
			OM09	39,7%
			OM10	4,9%
			OM11	3,7%
			OM12	3,6%
			OM13	22,5%
S. sahulensis	Australia		AUS01	3,7%
			AUS02	9,3%
S. teuszii	West Africa		WEA01	99,7%

Table 7.2 - Results from the BAYESCAN program showing all the 16 SNPS in common with the two first data sets that appear to be under directional selection. Between the two data sets, all 16 SNPs have similar values, showing positive alphas with values superior to 1 and FST ranging from 0.43 to 0.53.

Scaffold	Position	MAF 1%			MAF 2%		
		qval	alpha	FST	qval	alpha	FST
KB316843.1	125540	0.025436	1.3415	0.44567	0,035767	1.3356	0.50910
KB316843.1	7077457	0.027222	1.3756	0.45188	0,025365	1.4040	0.52121
KB316856.1	2313140	0.029163	1.3516	0.44731	0,04856	1.2445	0.49282
KB316870.1	15228638	0.018264	1.4600	0.46715	0,044396	1.2898	0.50102
KB316879.1	14241720	0.021101	1.4925	0.47332	0,037389	1.2860	0.50041
KB316886.1	524017	0.012914	1.4503	0.46572	0,031706	1.3624	0.51391
KB316887.1	11192557	0.0040508	1.5985	0.49279	0,038808	1.1749	0.47996
KB316888.1	10360926	0.023524	1.3658	0.44991	0,049458	1.2491	0.49350
KB316896.1	12639507	0.0064457	1.5751	0.48849	0,033896	1.2777	0.49891

KB316903.1	2029485	0.024280	1.4351	0.46271	0,04761	1.2870	0.50003
KB316906.1	6137432	0.030534	1.2723	0.43271	0,046576	1.2702	0.49704
KB316911.1	1802115	0.019305	1.4988	0.47431	0,027305	1.3292	0.50817
KB316928.1	1007044	0.020064	1.4530	0.46602	0,045539	1.2398	0.49192
KB316935.1	3593577	0.019527	1.3514	0.44730	0,016503	1.4161	0.52393
KB316992.1	3430274	0.014519	1.5293	0.47999	0,04027	1.2943	0.50169
KB317017.1	4070193	0.0073681	1.5285	0.47984	0,01807	1.4337	0.52740

Table 7.3 - Representation of the genes under directional selection with their corresponding acronyms, gene name and bibliography with information about each of the genes.

Gene	Gene name	Bibliography
D2HGDH	D-2-hydroxyglutarate dehydrogenase	(Han et al., 2018; Lin et al., 2015)
NEU2	Neuraminidase 2	(Koseki et al., 2012; Smutova et al., 2014)
NGEF	Neuronal guanine nucleotide exchange factor	(Chang et al., 2018; Rodrigues et al., 2000; Shi et al., 2010)
GLRX3	Glutaredoxin-3	(He et al., 2016; Pham et al., 2015, 2016)
PAPPA	Pappalysin-1	(Dunne et al., 2017; Petry et al., 2017)
TNC	Tenascin-C	(Mi et al., 2016; Rzechonek et al., 2018; Valdivieso et al., 2017)
MSRA	Methionine sulfoxide reductase	(Hu and El Haj, 2013; Noh et al., 2017; Singh et al., 2017)
CCT6B	Chaperonin containing TCP1 subunit 6B	(Agarwal et al., 2016; Ozaki et al., 1996)
FAT3	Protocadherin Fat 3	(Cheng et al., 2016; Krol et al., 2016; Mitsui et al., 2002)
SLC36A4	Solute carrier family 36 member 4	(Pillai and Meredith, 2011)
GALNT15	Polypeptide N-acetylgalactosaminyltransferase 15	(Cheng et al., 2004)
TMPRSS5	Transmembrane protease serine 5	(Mitsui, 2008; Yamaguchi et al., 2002)
DRD2	D (2) dopamine receptor isoform X2	(Binda et al., 2002; Gardner et al., 2008; Obregón et al., 2017)
GRM7	Metabotropic glutamate receptor 7	(Niu et al., 2015; Noroozi et al., 2016)
TSEN15	tRNA-splicing endonuclease subunit Sen15	(Budde *et al.* 2008; Breuss *et al.* 2016)
CNTNAP5	Contactin associated protein like 5	(Traut et al., 2006; Yu et al., 2018)
ANO10	Anoctamin-10	(Pedemonte and Galietta, 2014)
KBTBD11	Kelch repeat and BTB domain-containing protein 11	(Gong et al., 2018)
CUNH1orf21	Uncharacterized protein C1orf21 homolog	-
TMEM132E	Transmembrane protein 132E isoform X2	(Li et al., 2015b)
KIF13B	Kinesin-like protein	(Yamada et al., 2017)
TCERG1L	Transcript elongation regulator 1-like protein	(Miura et al., 2018)
EDEM1	ER degradation-enhancing alpha-mannosidase-like protein 1	(Lin et al., 2018)
FBXL7	F-box and leucine rich repeat protein 7	(Liu et al., 2015)

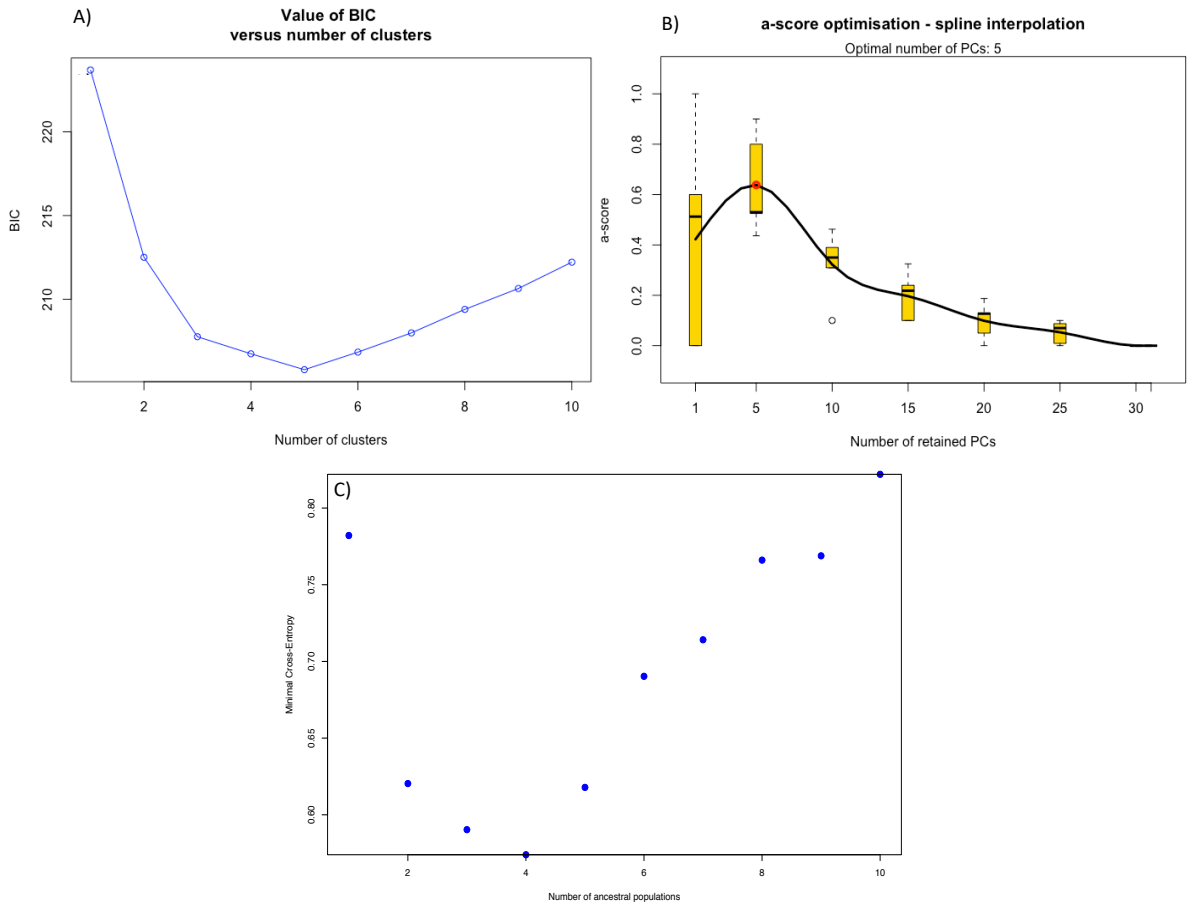

Figure 7.1 - The optimization of the number of clusters was a necessary step for both DAPC and SNMF programs. For DAPC analysis A) the value of BIC shows the number of suitable clusters for the data set in study. The optimal five clusters were obtained according to B) the number of retained Principal Components (PCs) each the optimal value was 5 PCs to retained in the analysis. A different value of a-score optimization affects the number of clusters obtained in BIC. For SNMF analysis C) the optimal number of clusters were obtained according to the Minimal Cross-Entropy. The lowest value was considered the optimal cluster for the data set analyzed in both programs.

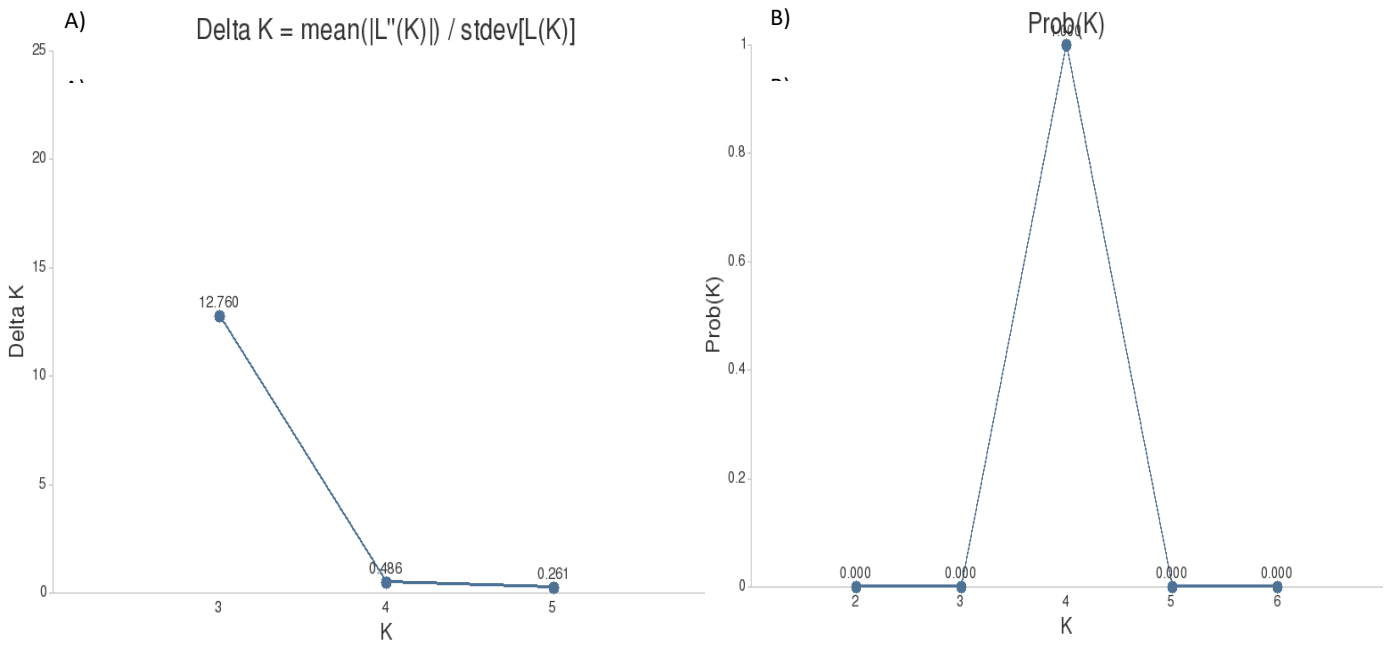

Figure 7.2 - In this study, the best value of K for STRUCTURE was determined based in two approaches: A) AK statistic by Evanno and B) *Ln(Pr(X|K))* by Pritchard. The highest value for both the approaches corresponds the ideal K for the analyzed data.

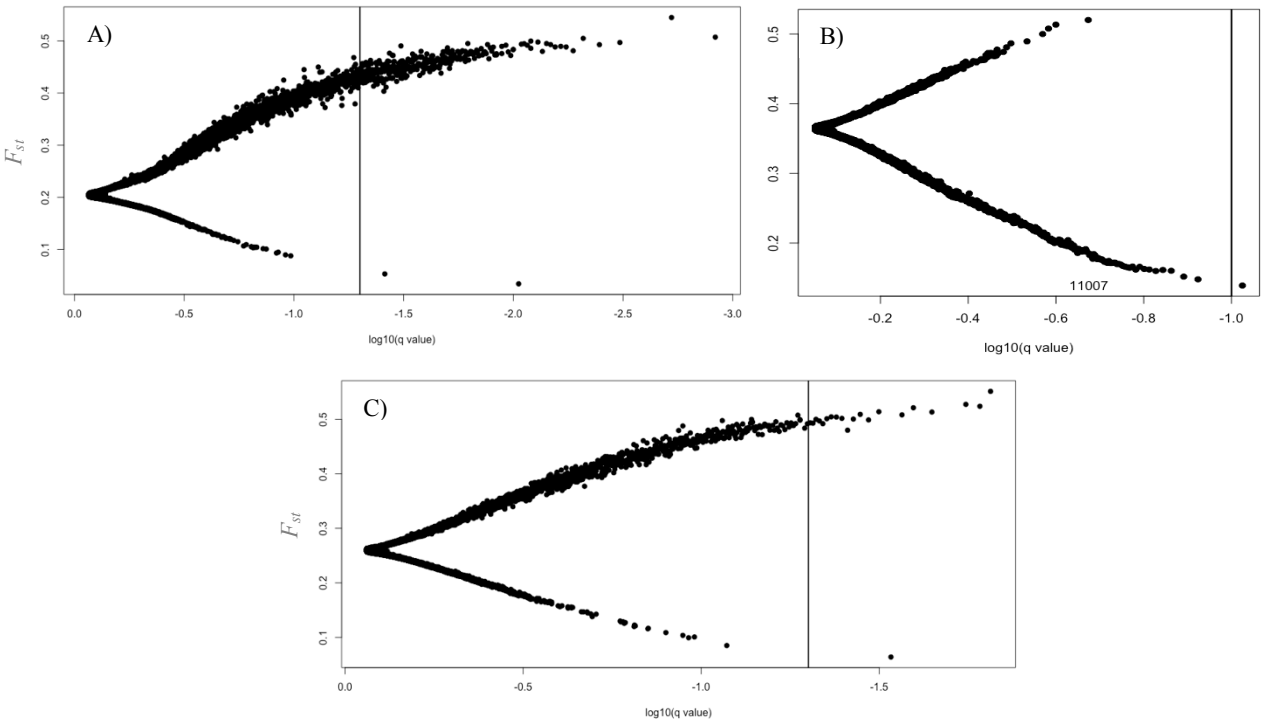

Figure 7.3 - Graphic representation of the results from BAYESCAN. All the three graphics show the FST distribution of the SNPs analyzed along the *qvalue* logarithm, and each one corresponds to different data sets with different MAF values A) MAF 1%; B) MAF 5% and C) MAF 2%. The line in all of the graphics corresponds to the FDR of 5% used to obtained the SNPS under directional selection. All the SNPs found on the right side of the line were under directional selection.

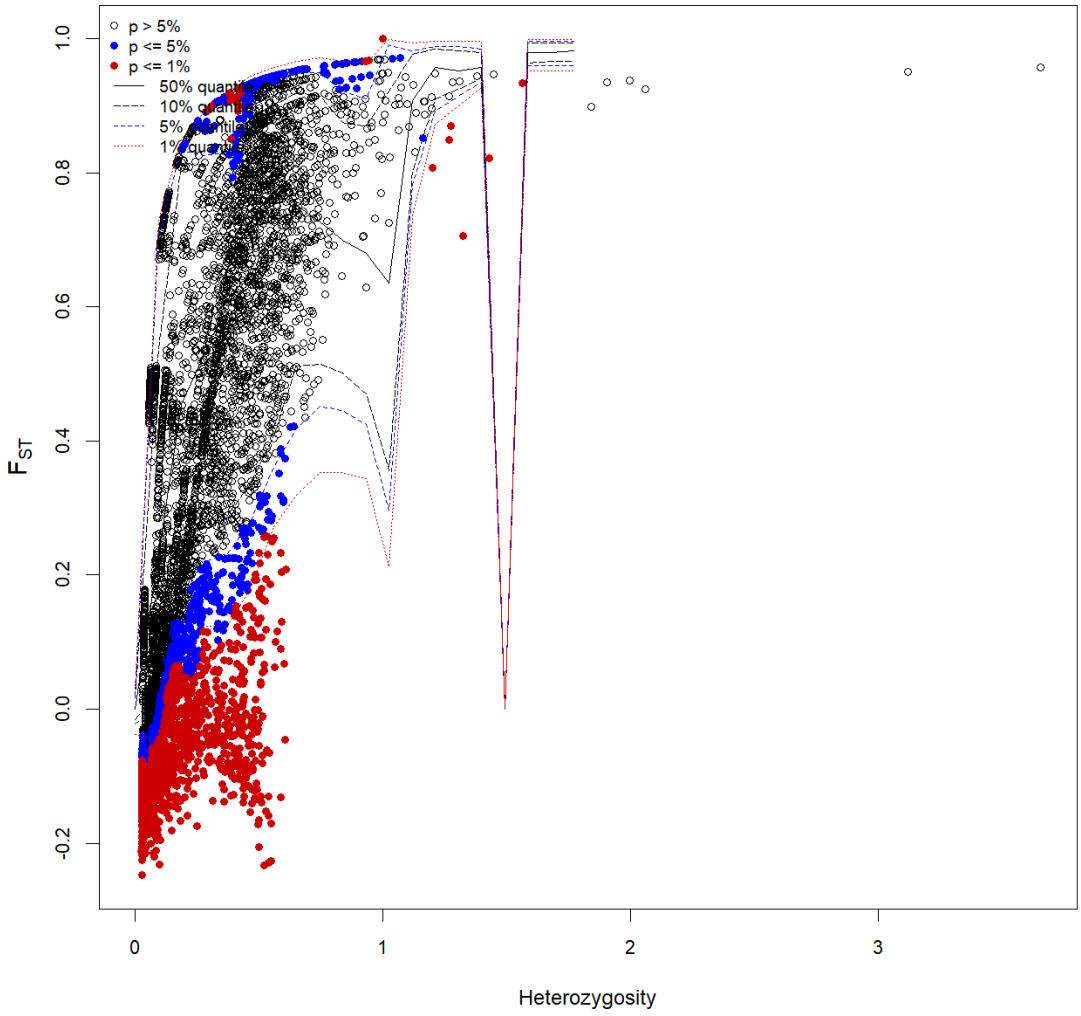

Figure 7.4 – Graphic representation of the detection of loci under selection from genome-scans based of Fst. Calculations done through Arlequin and graphic representation obtained from RSTUDIO.