PhD Thesis | George Pacheco

Flying Rats Shall No Longer Be
A Population Genomic Analysis of Fancy and Feral Pigeons

Advisor: Prof. Thomas Gilbert
Co-Advisor: Dr. Filipe Vieira
Internal Chairperson: Associate Prof. Rute da Fonseca
External Censor I: Associate Prof. Sanne Boessenkool | University of Oslo
External Censor II: Prof. Jon Slate | University of Sheffield
Submitted: 01 April 2019
Flying Rats Shall No Longer Be
A Population Genomic Analysis of Fancy and Feral Pigeons
To José Pachêco da Silva—a man I never met in life, yet who deeply influenced my path in it.
Existirmos—a que será que se destina?

_Cajuína_ | Caetano Veloso
Abstract

It has been supported by archaeological evidences that the rock pigeon (Columba livia Gmelin, 1789) was first domesticated during the Neolithic period in the Middle East, being initially used as a source of both food and fertiliser as well as in religious rituals. Later on, not only was the extent of its service to humankind spun into a wider variety of practical roles, but during the eighteenth century a great interest arose in the development and maintenance of so-called fancy pigeon breeds. As a result, over 230 pigeon breeds are officially recognised at the present, which encompass a well-known and truly fabulous range of phenotypic diversity. Furthermore, the long history of the pigeon’s domestication has also been tightly coupled with a long-standing process of feralisation, as domestic pigeons escaping from captive stocks successfully populated surrounding urban areas. In this way, populations of feral pigeons are found today to inhabit a great portion of the modern urban landscapes worldwide, leading many to view the feral pigeon as a pest species that requires active management. Moreover, as has been reported for wild ancestors of other domestic species, it is reasonable to suspect that the preservation of the gene pools of potential remaining wild rock pigeon colonies might be at risk due to admixture with feral pigeons. Although some studies have investigated the question concerning the complex phylogenetic relationships amongst the purebred pigeon breeds, these studies either included few breeds and few individuals per breed, or few genetic markers. On the other hand, no previous genomic work has properly addressed the natural history of feral pigeons so far.

During the last few years, reduced-representation library sequencing methods have been successfully established, which have facilitated the ready genotyping of thousands of genomic markers covering different regions of a species’ genome. Both the studies presented in this PhD thesis take advantage of this sequencing method, which allowed us to conduct the largest genomic studies to date on both purebred and free-living pigeon individuals:

In Chapter I, based on a genomic dataset including over 200 purebred pigeon individuals representing 67 breeds, we perform analyses of phylogenomics, MDS, Admixture, as well as we estimate several population genetic parameters. We report considerable genetic structure across the different pigeon breeds as well as substantial variation in their levels of genetic diversity. Importantly, we identify previously unreported interbreed admixture events. Moreover, we also investigate the suitability of genomic data generated through reduced-representation library protocols for studies aiming to correlate phenotypic variation with genomic modification.

In Chapter II, based on a genomic dataset including 450 free-living pigeons from 41 worldwide localities, we conduct phylogenomic, MDS and Admixture analyses as well as estimate several population genetic parameters aiming to shed first light on the pigeon feralisation question. Our results show that there is strong population structure across global populations of feral pigeons, and that most populations present signs of admixture with domestic lineages in considerably different degrees. Nonetheless, our results also indicate that some remote localities might still harbour wild colonies of rock pigeons that have been virtually free of contamination with domestic genotypes.

I løbet af de senere år er de såkaldte “Netsat Repræsentering Bibliotek (NRB)” sekventeringsmetoder blevet meget succesrige, hvilket har medført at det er blevet muligt at aflæse genotyper af tusindvis af genomiske markører fra mange dele af en arts genom. Begge studier i denne PhD afhandling udnytter denne sekventeringsmetode, hvilket gjorde det muligt at udføre det største genomiske studie til dags dato af både racerne og forvildede individer af tamduer:

I Kapitel I udfører vi, baseret på et datasæt af over 200 racerene duer fra 67 forskellige racer, fylogenetiske, MDS og Admixture analyser og estimerer flere populationsgenetiske parametre. Vi finder en del genetisk struktur henover de forskellige dueracer, såvel som stor variation i deres niveau af genetisk diversitet. Først og fremmest fandt vi tilfælde af blanding mellem forskellige dueracer. Herudover undersøger vi i hvor høj grad det er passende at anvende genomisk data genereret med NRB metoder i studier der ønsker at se på sammenhængen mellem fænotypiske træk og genomiske modificeringer.

I Kapitel II, som er baseret på et genomisk datasæt indeholdende 450 vildtlevende duer fra 41 lokaliteter spredt ud over hele verden, udfører vi genomiske, MDS og Admixture analyser og estimerer flere populationsgenetiske parametre med det mål for øje, at belyse historien om genforvildingen af tamduer. Vores resultater viser en kraftig global populationsstruktur mellem forskellige bestande af forvildede duer, og at de fleste bestande viser tegn på sammenblanding med tamme raser i meget varierende grad. Dog tyder vores resultater også på, at visse afsides lokaliteter stadig huser wilde kolonier af klippeduer som tilsyneladende er gået helt fri af introducerede tamdu genotyper.

Translated into Danish by Inger Eleanor Winkelmann.
Preface

Being myself from a rather biodiverse land and having been interested in nature since a young age, I could as a kid easily name by heart several snakes and spiders inhabiting my region that could cause great harm to any unadvised living being. As for the sea instead, I was also aware that many would be the shark species that could threaten almost any aquatic creature in those waters. Today, I find it interesting that although I was raised surrounded by splendor nature of delicate beauty, as a child I was always mostly impressed and attracted by those animals that most people would consider dangerous and frightening. Later on, when as a teenager I became very interested in dog breeds, I would by far favour the feared guardian breeds, such as the Rottweiler, the Dobermann, the Cane Corso or the Boerboel. Anyhow, for some reason unknown to me even today, my juvenile mind always demonstrated an admiration for those creatures that evolution had endowed with dexterous predatory or offensive skills.

Against this background, one might imagine how bemused and even disappointed I felt when I read Tom's reply to my email through which I had queried him as to whether he would have any interested in a project concerning animal domestication. It is true that his prompt reply demonstrated a great deal of enthusiasm regarding the subject to be addressed per se, however the organism he so eagerly suggested to be studied was—to say the least—an odd choice to me. As the content of my email sent to Tom might have indicated, I had come to Denmark pursuing a chance to do research in the field of domestication considering I intended to explore the question of how—through the process of artificial selection—wild and ferocious animals could have been turned into tame and docile ones. To that end, I hoped to genomically compare a given domestic animal against its wild ancestor. Therefore, I truly expected to work on a domesticated that would have an undeniably aggressive wild ancestor, such as the dogs of my childhood or even pigs. Nonetheless, Tom's reply mentioned pigeons. Pigeons?

I must confess though that by the time I first read his proposition I knew almost nothing about pigeons. I believe that I simply shared the common idea that pigeons were no more than filthy flying creatures—barely birds and carriers of a handful of diseases—that plagued cities across the globe. Thus, I could not really grasp at first the proposed connection between pigeons and domestication, given my certainty that pigeons were a mere yet rather troublesome invasive species. Even though greatly disconcerted by that initial reply,—for in my mind pigeons would not be domestic animals let alone have an aggressive wild ancestor whose genomic information could be used in my planned genomic comparison—I decided that some quick research into the pigeon subject would still be of good practice. It did not take long before I realised that the crystalised impression I hitherto had of pigeons was by far not consistent with reality. Not only had pigeons been domesticated, but also the amount of biological variation encompassed by pigeon breeds could easily rival those seen among the breeds of any other domestic animal. Furthermore, I also learnt through my preliminary research that the evolutionary history behind the so-called street pigeons was much more complex than those that are characteristic of invasive species, for these were actually somehow descendants of domestic pigeons—hence feral birds. Taking into consideration that these feral pigeons were virtually ubiquitous worldwide, one could instantly raise the relevant question: would the ancestor of the domestic pigeon, namely the rock pigeon, still to be found in nature? Profoundly surprised by my long-standing lack of knowledge of the matter that had
just been made evident, I decided to deepen my research. Soon after, I realised how terribly I had neglected the domestic pigeon,—to be probably attributed to my above-mentioned, almost complete disinterest in inoffensive beings during my youth—for the subject of domestic pigeons had in fact been used as a flag bearer for artificial selection in the On The Origin of Species—a title that I read during my teens.

Despite the fact that the domestic pigeon would not allow me to attempt to conduct my original research plan, I was sufficiently enchanted by its intricate evolutionary history. Thus, after digesting all that newly-acquired information, I sent a positive reply to Tom attesting my willingness to embark on the Pigeon Project. Several years have now passed since those first exchanged emails, and with the resulting benefit of hindsight I cannot avoid contemplating how the rock pigeon—including all its biological variants—has assisted me in my pursuit of a wider and deeper knowledge of evolutionary biology. To the extent that was I in the position to elect myself a species which could be used as a flagship for the modern field of Evolutionary Genomics, I would perhaps favour the rock pigeon; for be it subjected to the inventiveness of egocentric breeders, be it facing the difficulties and particulars of an unnatural urban life, or be it dangerously influenced by extraneous genotypes, the genome of this species has been evolving in many divergent ways.

George Pacheco
March 2019, Copenhagen—Denmark.
Acknowledgements

I believe that whenever nearing the end of any journey, the traveler naturally starts reflecting upon the paths that led to that specific termination. It was not different here. During the last months of my doctoral studies, I could not refrain myself from reflecting upon the entire voyage, especially endeavouring to identify its singular turning points. Naturally, this mental trip of mine culminated in my realization that I shall be in deep gratitude to both people and institutions which had profound influence on different periods of my life—from the development of my innate scientific curiosity to my ongoing academic training.

To my uncle Paulo Pacheco da Silva for introducing me to A Ciência da Abelha (The Bee’s Science) and for explaining me the roots of the word Universidade while he appreciated the taste of the ordinary Cachaça for the last time that day. I would realize years later that Science could be seen in every nook of this Multiverse and not only in the well-constructed hives of busy bees. With the passage of time, I would also inevitably acknowledge that the Universidade would be a much more complex and even malevolent entity than the one romantically pictured by you that night. Nonetheless, I will never forget our many Sunday evenings spent watching nature documentaries while we were comfortable in our swinging hammocks, nor your many failed attempts to persuade my mom that I was old enough to have a lizard or a tarantula as a pet.

To the Federal University of Rio Grande do Norte for providing me with the required basis—my forever Home University.

To Professor Lucymara Fassarella Agnez-Lima. You were my very first academic advisor and among many lab-related indispensable practices I shall never forget the most valuable lesson you taught me—that Science, as any other human activity, comes from within.

To Professor Wagner Franco Molina. You were my second academic advisor and indirectly taught me that a scientist must always experience an insatiable craving for reading. I cannot recall a single day when I entered your small and rectangular office and you were not immersed in some sort of scientific literature.

To the Danish Ministry of Science, Innovation and Higher Education for granting me with a Danish Government MSc Scholarship that concreted my will to study abroad. I shall be forever in great debt with Denmark.

To the University of Copenhagen for extending my basis—my forever Second Home University.

To the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico for providing me with a 4-year PhD Scholarship through the Science without Borders Programme, which made feasible my overseas doctoral studies.

To the Natural History Museum of Denmark for giving me the chance to realize my childhood dream of being part of such sort of human institution.

To the entire Section for Evolutionary Genomics for being such an easy-going group! Special thanks to Inger Eleanor Winkelmann and Ida Bærholm Schnell Dresen for their words of encouragement towards my intentions to make first contact with Tom; to Shyam Gopalakrishnan for trying to share with me a morsel of his knowledge on population genetics and bioinformatics; to Sarah Mak and Anna Fotakis for sharing their wonderful weeks of PhD thesis writing with me; also to Ângela Ribeiro for struggling to teach me something about the Avian World and for greatly contributing to the permanence of a minimal flame of classic thinking.
in my scientific soul.

To the **Bird 10K Project Consortium**, Mikkel Skovrind, Nerivania Godeiro, Andy Foote, Lisandra Zepeda, Marta Maria Ciucani, Mikkel Sinding, Morten Allentoft, Emily Cavill and Manuel Escalona for giving me the opening to significantly enlarge my doctoral experience.

To Jose Samaniego who I am confident that will never cease to regret of his decision to be the flatmate of a helpless biologist like the author of this thesis. I only hope that my level of annoyance was somehow bearable and that I did not make what was supposed to be your relaxing time at home a real purgatory of endless and trivial bioinformatics questions.

To what I call the **Little Portuguese Gang**: Paula Campos, Rute Fonseca, Gonçalo Themudo, Mafalda Costa and Ana Carolina for their friendship and for providing me with the unique opportunity to basically relearn my native language.

To Janina Kypke for being such a good friend throughout these PhD years of mine. I only hope that I managed somehow to reciprocate the great support you gave me.

To Professor Michael Shapiro and his team at the **University of Utah** for hosting me at their lab during my change of scientific environment. I am especially thankful to Anna Vickrey and Eric Domyan for all their lab support.

To my PhD advisor Professor Thomas Gilbert. I am well aware of how pigeons are close to your heart. I can only be very grateful that you entrusted this beloved project of yours to me. Thanks very much for all these years of scientific collaboration. At the end of this pilgrimage, it is quite easy for me to recognise that I did learn a great deal from you, and I am deeply grateful for all lessons.

To my family down in Brazil which has been always supportive of my desired exile, especially to my sister Daniela Pacheco who has lived most of her adult life away from the close attention of her elder brother. I shall also express my most profound and sincere gratitude to my parents Jorge Pacheco and Enauda Pacheco, who throughout their lives indeed invested everything they had but most importantly everything they never had towards my education.

To the family Radosavljevic Barros which assisted me throughout my stay here in Denmark, and provided me with a peaceful accommodation during my PhD thesis writing retreat. I am specially in debit with Link, for being himself an intercross of American Staffordshire Terrier and French Bulldog, was able to remind me of the powers of artificial selection by showing me the strength of his bite in every single morning.

To whom I deem be the most important person assisting my internal deconstruction seeking to convert a romantic scientist to an aspirant of modern researcher. Throughout this process Filipe Vieira was much more than a mere co-advisor. Ultimately, you were a friend—a generous friend.
List of Figures

Gutter Paradise #2 Portrait | Painted by Adele Renault ................................................................. Cover

Silhouettes of Animals and Plans | Images Under Licenses .............................................................. 15

Gnome - Genome Cartoon | Drawn by Tom Gauld ...................................................................... 17

Silhouettes of Animals and Plans | Images Under Licenses .............................................................. 20

Armando the Pigeon | BBC News .................................................................................. 21

Flying Flock of Feral Pigeons | Photo by Julia-Isabelle Gerber ......................................................... 22

Faroese Cliffs | Photo by Andreas Altenburger ............................................................... 23

Solitary Rock Pigeon | Photo by Andreas Altenburger ............................................................... 24

Pigeon Poker Face Portrait | Painted by Adele Renault .................................................................. 31

Camp’s Pretty Profile Portrait | Painted by Adele Renault ................................................................ 83

Camp Took a Bath Portrait | Painted by Adele Renault ................................................................ 131

Wet Specimen of Seira sp. | Photo by Nikolas Goia Cipola ................................................................ 133

Ide Specimen .............................................................................................................. 133, 137

Bat | Photo by Vitória Fernandes ............................................................................. 133, 149

Surfing Killer Whale | Photo by Andrew Foote ...................................................................... 133, 161

Black Mamba | Photo Under Public Domain .................................................................. 133

Moleques do Sul Cavy | Photo by Manuel Escalona .................................................................. 133

Seychelles Magpie-Robin Specimen | Photo Under License (Modified) ................................................................ 133

Goshawk Specimen | Photo by the Author ............................................................................. 133

Black Cuckooshrike | Photo by the Author ............................................................................. 133
# Table of Contents

## Introduction ........................................................................................................................... 13

*Every Cliché Holds a Great Deal of Truth* .................................................................................. 13

*The Reasons for Change* ........................................................................................................... 16

*The Rather Reduced Recipe That Still Makes a Good Dish* ....................................................... 18

*The Bad Editor Who Cannot Really Be Blamed* ......................................................................... 19

*The Trip Back or Almost* ........................................................................................................... 20

*Those Gone and The Well-Hidden Survivors* ............................................................................. 22

*References* ............................................................................................................................... 25

## Aims, Contributions and Structure ........................................................................................... 29

*Chapter I* .................................................................................................................................. 31

*Chapter II* .................................................................................................................................. 83

*Postface* ................................................................................................................................... 129

*The Heavily Bearded Man Was Seldom Wrong* .......................................................................... 129

*And The Pigeons Must Remain Aloft* ......................................................................................... 130

## Parallel Collaborations ............................................................................................................ 133

*Appendix I* ............................................................................................................................... 137

*Appendix II* ............................................................................................................................. 149

*Appendix III* ............................................................................................................................. 161
Introduction

Every Cliché Holds a Great Deal of Truth


***

Having been evolving on this rocky planet for a span of time that nearly matches the age of the planet itself\(^1,2\), Earth-like life has repeatedly and indefatigably generated biological diversity throughout aeons of cosmic evolution. Nonetheless, considering that the Earth has been subjected to several mass extinction events that wiped out of existence a great portion of the extant species during their respective times of occurrence\(^3\), Earth’s present-day biological diversity depicts a mere illustration of the organic plurality that has been forged on this astronomical body. In spite of this, the astonishing amount of biological variation that we currently witness on the planet is sufficient to demonstrate that the list of developmental opportunities that have been invented by life hereabouts is incredibly vast.

With so many options on the biological table, the developmental course of any living being on the planet ought to be strictly restricted. Thus, at any given time across a specimen’s development, crucial decisions concerning the activation or non-activation of certain developmental paths need be made; for perhaps its skin shall be almost translucent instead of completely opaque, its muzzle shall be elongated rather than nearly flat, its brain shall be heavily packed with tiny neurons and not loosely filled with large ones. Even after the newly-generated organism is fully formed, a rigid organic plan must be continuously followed, for perhaps when the right season comes, the specimen in question grows a giant flower instead of one of ordinary size, migrates to the west and not to the north, or favours the practice of sex over that of violence when striving to ameliorate tensions with conspecifics. In this way, it is of utmost importance that the specimen’s organic machinery is kept chained to the organic rhythm that is particular to it—until its very last breath of life. No deviation from its primordial design shall be ventured let alone permitted.

Against this background, life was faced with the challenge to create a system of encryption which could encode the entire set of organic instructions for the fabrication of new individuals. This system needed to be sufficiently stable throughout the period of the individual’s life, so as not to allow considerable erasing of the information present, or addition of allochthon information. Making use of an alphabet consisting of only four distinct characters (namely, A, T, C, G), life achieved the miracle of encapsulating such instructions in what
could be deemed proper biological novels. Thus, in the same way that a classic romance elegantly describes its several characters, giving them unique personalities and layers, placing them in particular ambiences and situations, these organic books dictate every movement of the organic gears. Also as with literature novels, the organic stories that are told by these biological novels are divided into well-structured chapters, which we define as chromosomes. Furthermore, just as acclaimed titles of the literature such *War and Peace* or *Don Quixote* cannot be fully appreciated by the reading of only one or a few loose fragments of their content, new individuals can only be assembled when their respective biological books are fully decoded. To these complex books of high biological literature which orchestrate the making of all the uncountable living beings that have existed on this planet, we give the name of *genomes*.

Despite the fact that these books of life are written in a much simpler language than most human languages, no human was formally educated in it. Therefore, we have been learning how to read part of genomes or entire genomes the hard way, through an act which we call sequencing. This important technical accomplishment has taken humankind considerable time and effort. This fact explains why even though the overall ways through which genomes are recorded has been known since the 1960s\(^4\), the very first genome from a free-living organism was only fully sequenced in 1995\(^5\), while the first genome of an eukaryotic organism was sequenced a year later\(^6\).

Sequencing can be performed in a range of different ways using different technologies, which are continuously being improved and developed. However, common to all sequencing approaches (at least until the present) is the fact that we never sequence each chromosome entirely in one go, but rather in short subfragments. Although the lengths of the fragments that can be sequenced varies considerably depending on the sequencing technology used, these fragments are almost invariably much shorter than the average chromosomal length (especially of more complex organisms). In addition, unless steps are taken to specifically enrich certain parts of specific chromosomes, the sequencing is usually performed on a random sampling of chromosomal fragments derived from all the chromosomes of a given genome. As a result, a mandatory step of genome sequencing is the computational reconstruction of these randomly sampled chromosomal fragments (which we call *reads*) into larger fragments (which are called *contigs*), followed by the subsequent assignment of these contigs into even larger fragments called *scaffolds*. Depending on the quality of the genome sequencing performed, these scaffolds can be further sorted into chromosomes, which ultimately should approximate the biological chromosomes that they were derived from.

One could argue then, that our current methods for genome sequencing allow us to sequence genomes, albeit almost as indirectly. As an analogy, one could think of reading *Les Misérables* by first completely tearing up this masterpiece until it was reduced to an assemblage of small, shuffled and meaningless fragments that would be read, then attempt to virtually reconstruct the entire book based on what was read and only finally strive to make sense of it. Undoubtedly, a great extent of the beauty of this piece of literature would be lost in this process, but its fundamental principle would be ultimately preserved. In this way, although far from ideal or error free, our current methods have allowed us to produce an ever-expanding list of sequenced genomes from organisms spanning different branches of the tree of life—from a marine worm\(^7\) to an electric fish\(^8\), from a tropical tree\(^9\) to a red algae\(^10\) (Fig. 1).

Nonetheless, it is noteworthy that sequencing does not equal comprehension. Even though the list of sequenced genomes continuously increases, it will take us much longer to make good sense of all
the sequencing conducted. Similar to what is done in the field of \textit{Comparative Literature}, which contrasts different written works with regards to their literature genres, writing periods and writing styles, genomicists gain much knowledge by comparing the different genomes of divergent organic forms. These comparisons have the potential to reveal both fundamental genomic features as well as genomic characteristics that are particular to certain traits common to unrelated clades. Not surprisingly, this specific field is called \textit{Comparative Genomics}\textsuperscript{11} and it is in continued development as we sequence more and more genomes.

Although we lack a complete understanding of genomes, based on the comparisons among the genomes we have already sequenced and attempted to make sense of, we can outline some overall genomic

\textbf{Figure 1 | Illustration of Organisms with Read Genomes.} A) Capybara\textsuperscript{59} B) Giant Panda\textsuperscript{23} C) Stoat\textsuperscript{60} D) Golden Eagle\textsuperscript{61} E) Bonobo\textsuperscript{62} F) American Flamingo\textsuperscript{63} G) Pepper\textsuperscript{64} H) Emperor Penguin\textsuperscript{65} I) American Crow\textsuperscript{63} J) Zebrafish\textsuperscript{66} K) Common Ostrich\textsuperscript{63} L) Thale Cress\textsuperscript{66} M) Water Bear\textsuperscript{67} N) Whale Shark\textsuperscript{68} O) Hemp\textsuperscript{69}. 
trends. For instance, genome length (also referred to as genome size) does not correlate linearly with biological complexity or mass—indeed that an ordinary-looking herbaceous Japanese plant has a genome roughly seven times larger than the genome of a tall pine tree, and a salamander ranging in length of less than half a meter has a genome that is roughly ten times larger than that of a hominid. Furthermore, the field of Comparative Genomics has already uncovered particular genomic characteristics that underpin common traits that are present in diverging clades, such as extended longevity, mammalian adaptations to an aquatic life and obligate bamboo diet.

***

The knowledge that we have acquired from the few genomes we have already sequenced has been enough to teach us about the utmost relevance of this kind of biological literature. Therefore, we shall continue to expand our list of sequenced genomes, for by sequencing genomes we are broadening our understanding of Earth-like life itself. In addition, in the same way that a bibliophilic bookworm becomes better at comprehending and appreciating literature as she amasses experiences through reading a larger variety of literature titles, the best way for us to be in the position to grasp all the nuances of a given genome is by analysing it alongside other genomes.

**The Reasons for Change**

It goes without saying that basically all contemporary scholars working in the field of genomics vividly daydream about a time when an well-curated genome library will exist, which will represent the genomes of all the species on Earth. Nonetheless, this day has not arrived yet. On the contrary, this is far from reality at the present as we have sequenced just a small fraction of the entire genomic constellation that is out there to be sequenced. However, some rather ambitious projects have been commenced, which aim to considerably boost the numbers of genomes sequenced. There is, for instance, the Bat 1K Project, which aims to sequence the genomes of all living bat species. Similarly, with the goal to sequence the genomes of all living avian species, there is also the Bird 10K Project, a project that I myself have played a small, yet vital, role in during my doctoral time in Copenhagen. Placed above all these clade-focused projects, the Earth BioGenome Project intends to eventually sequence the genomes of no less than all Earth’s eukaryotic species. Indeed, all these projects and many others of a similar kind shall be deemed commendable (Fig. 2), albeit there are simply too many genomes to sequence. Thus, it will certainly take us decades before we build a genomic library that may rival the Great Library of Alexandria regarding completeness.

Being a direct product of genomes, biological diversity expands as genomes vary. Thus, Earth’s immeasurable biological diversity—that today delays the completion of our desired library of genomes—is the direct product of a long and continued history of genomic changes. Such correlated expansion is in fact, only made possible because, while the process of cross-generational genome copying is certainly far more precise than the copying executed by the most dedicated of the Gregorian amanuenses, it is far from perfect. In fact, had this copying system been immaculate, organic evolution as we know it could never have occurred. Ultimately, therefore, genomic differences accumulate through time, just as how ancient manuscripts that have been extensively copied throughout the centuries have accumulated misspellings.
Eventually, however, more dramatic textual changes occur. For example, sometimes a whole section of the genome is copied twice instead of singly, or not copied at all. It may also happen that some genomic sections are inverted while copied, or copied into genomic locations that are far from their original position. Perhaps the most extreme case of defective genomic copying is when the whole genome is copied not once, but twice (e.g. ref.²⁷,²⁸). Moreover, due to so-called epigenetic modifications, genomes are in fact written obeying a complex kind of accentuation system. Thus, in the same way we need to be attentive when copying any piece of French text as in this language *Mais* means *But*, while *Maïs* means *Maize*, genomes must be copied in a way that preserves their accentuation patterns, seeing that modifications in this delicate system may lead to important changes in how the genomes are decoded and interpreted by the biological machinery²⁹.

Altogether, these accumulated genomic differences at the intraspecific level may give rise to noticeable differences among the individuals of a given species. As a result, this intraspecific variation serves as a fundamental substrate for the evolutionary process by natural selection, as well as for other evolutionary mechanisms. This evolution may lead to local adaptation and even, if a sufficient level of genomic differences accumulates, the emergence of a new species. Naturally, given its importance for the process of evolution, genomicists have a deep interest in understanding how genomes vary intraspecifically, and this important subject is of particular concern to the specific field of *Population Genomics*³⁰. However, in order to properly address this question, population genomicists must typically sequence the genomes of several individuals belonging to several populations of given species.
The Rather Reduced Recipe That Still Makes a Good Dish

Recently, important improvements in sequencing technologies have occurred, which led to a considerable decrease in the costs involved in genome sequencing. However, this process is still quite costly. This inescapable reality prevents most of the researchers interested in genomics from both sequencing genomes of different species, as well as genomes from different populations of certain species. Under this circumstance, genomicists were forced to take the only possible path—they would sequence more by actually sequencing less. Given the high complexity of genomes, genomicists argued that perhaps by sequencing only fractions of the genomes, they might gain as much analytical power as if they had sequenced the genomes thoroughly. This concept is similar to that employed in the field of Comparative Literature, where a modern researcher may not need to meticulously analyse the whole titles of *Ensaio Sobre a Cegueira* by José Saramago (written in Portuguese), *Heart of Darkness* by Joseph Conrad (written in English) and *Cien Años de Soledad* by Gabriel García Márquez (written in Spanish) to come to the conclusion that, at least linguistically, the first has much more in common with the last than with the second. Of course, it is critical to remember the caveat that some few words or even expressions might not obey this general trend. Nonetheless, if the researcher makes sure that she has compared a minimal number of textual fragments, and that these were selected at random, she can be fairly confident that her conclusion would be the same had she analysed the entire textual contents of all three books.

Inspired by this general idea, genomicists have developed a method that allows them to sequenced randomly chosen portions of genomes, from many individuals in a way that the same genomic data is recovered for each individual. This method initially uses a restriction enzyme that cuts the genomes in question at a predetermined sequence (e.g. ATTATC). Consequently, several genomic fragments are generated through this first process. The method then limits the actual sequencing to be performed on a certain number of bases of these fragments that flank the located cut-sites. As the chances of finding the specific cutting sequencing in any given genome are much lower than if any pattern was accepted, and as this specific cutting sequencing will be randomly located across any genome, this method guarantees that only a fraction of each genome is actually sequenced and that the sequenced genomic contents were chosen at random.

Although there are several variations of this method, all the developed modifications are collectively referred to as Reduced-Representation Sequencing methods for the self-explanatory reason that this is precisely what these methods yields. Regardless of their specificities, these methods have successfully been used by genomicists studying a wide variety of organisms and evolutionary questions. For instance, genomicists have now used these methods to assess whether ecological divergence has intensified the levels of diversification of a plant complex that inhabits the Páramos, to infer the levels of genomic diversity and distribution of genomic variation in the Muskox, to test how many independently evolving lineages exist within a species complex of shrubs.

Indeed, while genomicists cannot fully live in the genomic paradise they all dream with, they found a way to keep themselves busy comparing and interpreting fractions of genomes; for even though their comparisons and interpretations are not based on the sequencing of complete genomes, genomicists can
indeed still draw important conclusions from these analyses. After all, studies based on some sequencing is undoubtedly better than studies based on no sequencing at all.

**The Bad Editor Who Cannot Really Be Blamed**

Although genomes undeniably change as time passes\(^{37}\), generally a long stretch of time is required before a genome has accumulated enough changes that might be translated into noticeable phenotypic modifications. Thus, it is only under very special circumstances that genomes, and hence species, can diversify in relatively much shorter periods of time. Some of these special conditions are natural (e.g. ref.\(^{38}\)), but perhaps the condition under which a species and its genome can most rapidly deviate from its ancestral state is directly conducted by humankind—the unique condition of artificial selection. This human-mediated evolutionary process has produced a myriad of variant forms out of several naturally-occurring organic layouts. For instance, one single plant species has been used as prima materia for the creation of a list of many crops, which, even though popular, would not be deemed to belong to the same species by popular sense. This list includes Cabbage, Cauliflower, Broccoli, Brussels Sprout, Kohlrabi and Kale\(^{39}\).

Nonetheless, despite the fact that domestic variants are often very dissimilar in comparison to their respective ancestral states, it is noteworthy that genomes do not need to be substantially changed before this result is achieved. Just as skillful writers such as Dostoevsky or Tolstoy are able to dexterously alter the direction of a story by meticulous modifications in this or that character or scene, artificial selection can produce marvelous results by leveraging relevant point mutations. For example, the genome of an Icelandic horse is nearly identical to the genome of any other horse breed, or even the genome of the wild horses that were ancestral of all domestic horses. As a consequence, despite the great deal of phenotypic variation within the numerous breeds of horse, given the close similarity of their genomes, all are considered to belong to the same equid species—*Equus ferus*. Not surprisingly, a modification of a single base in the horse genome has been shown to have a significant effect on the Icelandic horse’s unique gait pattern\(^{40}\). As a corollary to this observation, we are led to conclude that artificial selection is the art of blindly perusing minor genomic changes that radically cause major alterations in the phenotype expressed.

This particular way of evolution leaves unique genomic signatures on the genomes of domestic lineages, and genomicists soon realised that they could take advantage of these specific genomic scars produced by artificial selection in their attempt to identify phenotype-genotype correlations\(^{41}\). Therefore, not only has the last decades witnessed a great effort made by the genomicists’ community towards the sequencing of the genomes of several domestic organisms (Fig. 3), but also an increasing number of genomic investigations on domestic lineages aiming to decode the genomic basis of several morphological and behavioural traits seen in these lineages. Confirming the theoretical framework, these studies have been of paramount importance for both the expansion of our understanding of fundamental questions concerning molecular evolution as well as of the genomic features behind specific phenotypes. Given that, one would not be mistaken to attest that the development of the field of genomics as a whole is inextricably intertwined with the progress of our knowledge of domestic-animal genomics.

***
On May 3rd 1859, a Victorian editor addressed a fairly long letter to a publisher explaining why he discouraged the publication of a manuscript he had been asked to edit. The work in question attempted to expound on how new species could be originated. However, the editor considered that the book described an innovative theory without presenting the required proofs and evidences. Therefore, concurring with an alternative first proposed by a famous naturalist, the editor suggested that the author should rewrite his work so it mainly dealt with his observations on pigeons as these animals were “curious, ingenious, and valuable in the highest degree”. Moreover, he stressed, “Every body is interested in pigeons.” [sic] (Fig. 4). Nonetheless, even though admittedly not fully aware of all the organic mechanisms underpinning his own theory, the author, who had been working on his treatise for more than 20 years, was convinced enough of the value of this work as it was. Thus, he firmly refused to dramatically change his work except for a few minor edits. In the end, the publisher accepted his denial and few months later that year the manuscript was finally published.

What followed is history, but if only they could read genomicish.

**The Trip Back or Almost**

In June of 1949, George Orwell published his well-known dystopian novel entitled *Nineteen Eighty-Four*. It contained the following phrase:

*Left to themselves, like cattle turned loose upon the plains of Argentina, they [the proles] had reverted to a style of life that appeared to be natural to them, a sort of ancestral pattern.*

Similarly to Orwell’s imaginative proles and the real cattle of the Argentinean plains, domestic pigeons have also been left to strive for themselves (Fig. 5). No longer under the hefty influence of breeders, these...
abandoned pigeons have also themselves reverted to a state which somehow resembles that of free-living wild rock pigeons. Nonetheless, the urban environments which these birds inhabit are indeed quite disparate when compared to the cliffs and rock ledges that were once occupied by the ancestral rock pigeon. Furthermore, considering their domestic background, the genomes of these birds are assuredly laden with indelible genomic scars, which could potentially impede their full reversion to an absolute wild state. Consequently, it has been suggested that the feral pigeon lineage has been treading along its own evolutionary path, evolving under the unique pressures of urban life.

In the light of that, one might be inclined to deem the pigeon feralisation scenario a rather unique evolutionary process. However, one might not fail to acknowledge that besides the cattle in the Argentinean plains and the pigeons pecking in the streets of the modern world, many other domestic lineages have gone through a process of feralisation (e.g. the dromedary, the dog, the cat and the pig), and that these cases followed the same general characteristics of the pigeon case. Thus, one might indeed recognize that the evolutionary process known as feralisation is as common as it is unique when compared to the domestication
process, nay, perhaps feralisation is even more unique given that it has generally occurred in much shorter periods of time, which provide us the unique opportunity to study much rarer evolutionary cases of extremely recent occurrence. If these facts are to be acknowledged, one shall conclude that the process of feralisation ought to receive relevant scientific attention in the same way domestication has rightfully received during the last centuries.

Those Gone and The Well-Hidden Survivors

It is widely believed that through several Kafkaesque events of destruction, numerous titles that were only to be found in the Great Library of Alexandria were permanently lost. Thenceforth, humankind has been and forever will be precluded from regaining access to those ancient texts. Although there is ongoing academic dispute over the true nature of these events and over which contents were and were not lastingly lost, this popular myth of remarkable knowledge loss serves to constantly remind us of how fragile encoded information actually is.

As aforementioned, whilst species have been naturally and steadily erased from existence throughout the geological time, new species and hence new genomes have been concomitantly formed on the planet. Even though these processes of species’ annihilation and establishment generally follow a qualitative trend, where species are often deemed either extant or extinct, every now and then these processes may occur in much more subtle ways (e.g. ref. 50,51). Perhaps, the evolutionary scenario that can best exemplify this particular

Figure 5 | Flying Flock of Feral Pigeons at the Rynek Główny Square in Kraków, Poland.
situation is when we attempt to ponder whether the wild ancestors of some extant domestic lineages still exist or have been lost. This issue is especially raised when the domestication in question happened long ago and when humans successfully maintained the resulting domestic lineages across the natural habit of the corresponding wild ancestor; for, since the modifications caused by domestication seldom prevent domestic lineages from intercrossing with their respective wild ancestors, the possibility that current wild populations are in fact descendants of intercrosses between now-extinct wild ancestors and ancient domestic lineages cannot be entirely ruled out. In other words, it may well be that populations considered to safely guard the genomic information of wild ancestors of now-domestic lineages might in fact be already corrupted by insertions of genomic sections of domestic origin. Naturally though, this poses the question of how much admixed a wild population must be to no longer be deemed wild. As one might expect, rather gray zones are often reported. However, modern sequencing technologies have greatly assisted us in expanding our comprehension of these convoluted evolutionary encounters (e.g. ref.52–54).

In spite of these genomic particularities, the rock pigeon shall certainly be considered an extant species represented by both a countless number of purebred domestic pigeons, carefully watched by dedicated breeders worldwide, and an even larger quantity of free-living pigeons that populate urban landscapes across the globe. That wild rock pigeon that once inhabited the cliffs and rock ledges (Fig. 6) —that particular rock pigeon may well be forever gone. Forever gone alongside that unique genome. Nonetheless, perhaps perfect copies of those books thought to have been lost in Alexandria are still lying in some library yet unstudied by Archaeology; perhaps some immaculate colonies of wild rock pigeons yet to be discovered still inhabit remote and isolated locations (Fig. 7).
Figure 7 | Solitary rock pigeon individual in the Streymoy region on the Faroe Islands.
References

64. Kim, S. et al. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-
Aims, Contributions and Structure

The ultimate goal of the herein presented PhD thesis was to shed light on the genomic consequences and singularities of the rock pigeon’s evolutionary dynamic state. To this end, we took advantage of both its privileged genomic position—with a well-curated genome assembly and available datasets of fully-sequenced genomes—as well as the recent method of Genotyping-by-Sequencing.

Chapter I presents an article which tackles the long-standing question of the complex evolutionary relationships among pigeon breeds and investigates how artificial selection has been differently affecting the genomes of these many breeds. Building up on previous works, we conducted the largest study on this question to date. Our analyses produced both corroborating and conflicting results, which not only demonstrate the specific complexity of the pigeon case, but also support the general concept that the phylogenies of domestic breeds will inevitably be highly entangled due to the uniqueness of evolution by means of artificial selection. This article is entitled *Darwin’s Fancy Revised: An Improved Understanding of the Genomic Constitution of Pigeon Breeds* and is currently under review in the *Molecular Biology and Evolution* journal.

Chapter II presents a pioneering article which addresses the question of the pigeon feralisation and how this process has been potentially threatening the preservation of the wild rock pigeon. More specifically, this article investigates the evolutionary relationships among several populations of feral pigeons worldwide, and describes the genetic background of these populations. Importantly, the results presented in this article suggest that some pigeon populations have been less exposed to domestic and/or feral genotypes, which indicates that these populations might be of considerable relevance for the preservation of the rock pigeon as a wild species. This article is entitled *On the Origin and Spread of Feral Pigeons* and is currently in preparation to be submitted to the *Communications Biology* journal.

As the leading author of both these presented articles, I was responsible for the coordination of all the required procedures, such as sample collection, data generation, computational analyses and scientific writing.

Moreover, a list describing parallel projects in which I was involved during my doctoral studies is presented in the Parallel Collaborations section. Three published articles that resulted from these collaborations are presented in the three Appendices. Nonetheless, please notice that these parallel collaborations are presented here for illustrative purposes only. Therefore, the assessment of this PhD thesis shall not be performed based upon them.
Chapter I

The reasoning to generate the dataset analysed in this chapter was conceived by Prof. Gilbert and I so we could better investigate the pigeon feralization history as a direct continuation of my MSc studies. Nonetheless, as we started to comprehend the complexity of the pigeon’s history of artificial evolution, this dataset started to receive more and more attention until we finally decided that it should be scrutinized in a separate article, where we would have the opportunity to discuss our findings regarding the pigeon breeds in appropriate length. Due to the uniqueness of the analyses herein performed, I had to create together with Dr. Filipe Vieira a bioinformatic approach that could cope with this pioneering genomic experiment design. Moreover, this project led me to experience a change of scientific environment, for I performed the relevant lab work at the lab of Prof. Shapiro, which is based at the University of Utah in the USA. I strongly believe that this experience was of the most importance for my internationalized education.
**Darwin's Fancy Revised: An Improved Understanding of the Genomic Constitution of Pigeon Breeds**

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Molecular Biology and Evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>MBE-18-1053</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>13-Dec-2018</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Pacheco, George; University of Copenhagen, Natural History Museum of Denmark  
Van Grouw, Hein; Natural History Museum, Department of Life Sciences  
Shapiro, Mike; University of Utah, School of Biological Sciences  
Gilbert, Tom; University of Copenhagen, Natural History Museum of Denmark; Norwegian University of Science and Technology, NTNU  
University Museum  
Vieira, Filipe; University of Copenhagen, Natural History Museum of Denmark |
| Key Words: | Animal Domestication, Pigeon Breeds, Population Genomics, Genotyping-by-Sequencing |
Article I Discovery

Darwin’s Fancy Revised: An Improved Understanding of the Genomic Constitution of Pigeon Breeds

George Pacheco1*, Hein van Grouw2, Michael D. Shapiro3, M. Thomas P. Gilbert1,4 & Filipe Garrett Vieira1

*george.pacheco@snm.ku.dk

1Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark.
2Bird Group, Department of Life Sciences, Natural History Museum, Akeman Street, Tring, Herts. HP23 6AP, United Kingdom.
3School of Biological Sciences, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112, USA.
4NTNU University Museum, Norwegian University of Science and Technology, NO-7491, Trondheim, Norway.
Abstract

Through its long history of artificial selection, the rock pigeon (Columba livia Gmelin, 1789) was forged into a large number of domestic breeds. The incredible amount of phenotypic diversity exhibited in these breeds has long held the fascination of scholars, particularly those interested in biological inheritance and evolution. In order to broaden our understanding of the complex evolutionary history of pigeon breeds, we conducted the most extensive phylogenetic analysis of the group using over 200 domestic individuals representing 67 breeds, two feral pigeons and one outgroup. We find considerable structure and variation in the genetic diversity of the different breeds, and identify previously unreported interbreed admixture events. Finally, we investigate the suitability of genomic data generated through reduced-representation library protocols for studies aiming to correlate phenotypic variation with genomic modification.

Keywords: Animal Domestication, Pigeon Breeds, Population Genomics, Genotyping-by-Sequencing.
Chapter I

Introduction

Domestic animal lineages have long been appreciated for their value as model systems with which to identify the genomic mechanisms underlying (their often remarkable) phenotypic variation (Andersson and Georges 2004), thus contributing to our understanding of fundamental evolutionary processes (Andersson et al. 2012; Imsland et al. 2012; Rubin et al. 2012). In this regard, domestic pigeons exhibit some of the most remarkable biological variation, and as such attracted the interest of Charles Darwin himself. Not only did he, on November 4th, 1855, end a letter to his friend and colleague Charles Lyell, with the following words: “I will show you my pigeons! which is the greatest treat, in my opinion, which can be offered to human being [sic]” (Darwin 1855), but he also opted to introduce his theory of natural selection by discussing the role of artificial selection in the creation of pigeon breeds (Darwin 1859). Furthermore, in his later book that focused specifically on describing the products of both animal and plant domestication (Darwin 1968), two whole chapters were dedicated to pigeons, where he argued that despite their immense biological diversity, all breeds descended from a single species, the rock pigeon (Columba livia Gmelin, 1789) (Darwin 1968). It is unsurprising therefore that pigeons have also been of interest to geneticists since the field’s earliest days (Staples-Browne 1908; Bonhote and Smalley 1911), due not only to their aforementioned astonishing phenotypic diversity, but also the ease with which samples can be obtained from domestic stocks and cross-breding experiments undertaken.

The rock pigeon was first domesticated in the Mediterranean region as early as the Neolithic period (Johnston and Janiga 1995). This initial domestication effort probably focused on traits of direct benefit to humans, such as tameness or meat production. Once domesticated, pigeons spread with humans throughout Eurasia, and subsequently diversified under the influence of local needs, constraints and cultures. Much later, during the Victorian era, the outcomes of pigeon breeding shifted focus to the purpose of generating unique phenotypes, leading to the development of fancy breeds that exhibit a wealth of exuberant traits. Ultimately, the complex interaction of their geographic distribution, periods of creation, and purposes of selection gave rise to today’s extremely heterogeneous collection of pigeon breeds, which breeders have attempted to classify based on characteristics such as function, morphology, vocal abilities, and origin.

In this study, we aimed to leverage on population genomic techniques in order to improve our understanding of the complex evolutionary history of pigeon breeds, by complementing the results of previous analyses based on fewer genetic markers (Stringham et al. 2012) and breeds (Shapiro et al. 2013). Specifically, we generated a genome-wide dataset encompassing more than 200 domestic individuals representing 67 breeds, two feral pigeons and one outgroup, and used it to reconstruct the most inclusive phylogeny to date for the group. We report previously unknown admixture events that occurred between different branches, and genomically characterize several pigeon breeds, providing support for their classification into distinct groups.
Results & Discussion

Phylogenetic Relationships Among Pigeon Breeds

To broaden our understanding of the phylogenetic affinities among pigeon breeds, we conducted a maximum-likelihood (ML) phylogenetic analysis for all samples that passed our quality controls, using *Columba rupestris* as outgroup. Twenty-three of these samples were sequenced using both GBS (here) and WGS (previously (Shapiro et al. 2013)) in order to control for whether any significant bias might be introduced by our joint analyses of these two types of genomic data. Through the merging these GBS samples with their respective WGS samples, we also created combined samples (WGS-GBS), totaling 23 triplicates. This dataset (Dataset 1; 184 GBS, 50 WGS and 23 WGS-GBS) comprised 1,997,420 total sites (including monomorphic) with coverages at potential GBS loci ranging from 7.03X to 323.05X (mean 85.22X), and missing data from 0% to 1.55% (mean 0.16%) (Supplementary Spreadsheet). To check whether the GBS method produced data at locations spanning the entire pigeon genome, we performed a regression between the size of each scaffold and the number of sites reported (Supplementary Fig. 7a), indicating that the GBS protocol successfully yielded sites randomly distributed across the pigeon genome.

Our phylogeny shows that all triplicates cluster together with 100% bootstrap support, confirming the absence of bias when either analysing GBS and WGS data under a single pipeline, or when merging these two data types (Fig. 1). We noted, however, that the GBS libraries systematically presented longer branch lengths with respect to their replicates. This hints at an excess of low frequency SNPs (e.g. singletons and doubletons) in the GBS and WGS-GBS samples. Since these are not shared with the WGS replicate, we conclude they are likely sequencing errors that were considered as alternative alleles (due to the impossibility of removing PCR duplicates). Thus, we believe that the use of a reduced-representation library sequencing method that allows for the removal of PCR duplicates (e.g. paired-end RADseq) could considerably reduce this methodological issue.

We used the current classification of the American National Pigeon Association (NPA) as a reference for pigeon breed grouping. We did so for the sake of convenience, and acknowledge that some of these groupings are based on historical reasons and do not necessarily reflect phylogenetic relationships. The NPA classifies pigeon breeds into 9 groups, based largely on morphological and behavioural traits: Form; Wattle; Croppers & Pouters; Color; Owls & Frills; Trumpeters; Tumblers, Rollers & High Flyers (TRHF); Structure; and Syrian (see Methods for details). Overall, the topology of our ML phylogeny is consistent with previous analyses of both WGS (Shapiro et al. 2013) and microsatellite (Stringham et al. 2012) datasets, and successfully recapitulates the seven principal clades described in the latter, while also highlighting that some of the NPA groups are not monophyletic.

Our phylogeny did, however, highlight several differences concerning the topological placement of some breeds. For instance, the previous analysis of WGS (Shapiro et al. 2013) placed the Jacobin as sister to the Danish Tumbler (together with the remaining TRHF), while we find that the Jacobin forms a clade with the Old Dutch Capuchine, which is a sister group to the Trumpeters. This Jacobin-Old Dutch Capuchine relationship has been previously reported (Stringham et al. 2012), and is consistent with their morphological resemblance (e.g., both have a hood).
and shared ancestry (Moore 1735; Levi 1986; Stringham et al. 2012). We hypothesise that the affinity of the Jacobin to Tumblers and Trumpeters, may be due to shared genetic background between the three breeds. Another difference is that while one previous study (Shapiro et al. 2013) placed the Carneau in the clade of Pouters and other large-bodied breeds such as the Runt and King, our phylogeny places it as sister to the Scandaroon, and sharing common ancestry with the Homers and the English Carrier. Originally the Carneau was bred in France for meat production (Levi 1986), thus it seems logical that it could have been created out of larger breeds, such as the early archetype Carrier (also known as Bagadet) (Moore 1735), which was probably an ancestor of the English Carrier, Scandaroon and Racing Homers. The French Carneau was later imported to the United States around 1900, where its appearance has been dramatically modified through outcrossing with other breeds. Thus, the modern day American Carneau is sufficiently different from its European ancestor, that it can almost be considered a different breed (Levi 1986). A third notable difference is how the Laugher clusters with Croppers & Pouters in our phylogeny, while it is found as a sister breed to the Fantails and some TRHF in one previous phylogeny (Shapiro et al. 2013).

We also noted how several pigeons labelled as belonging to a single breed were found on different branches of our phylogeny. For example, the two Mindian Fantail (a breed not recognised by the NPA) samples did not form a monophyletic group. One sample is an outgroup to all Indian Fantails, while the other is an outgroup to all Fantails and Indian Fantails. We note that the Mindian Fantail is the product of a recent outcross, created with the goal of miniaturising the Indian Fantail. In order to achieve this result, breeders outcrossed the Indian Fantail with other breeds (namely small Tumblers; D. Skiles, pers. comm. with MDS), which could explain these phylogenetic incongruences. Similarly, a recent study found that dog breeds under development would also have a tendency to not form monophyletic clades (Parker et al. 2017). The two samples belonging to the Iranian Tumbler breed (a breed also not recognised by the NPA) did not cluster together either. This may indicate that one of these samples was recently outcrossed or erroneously labeled.

Overall, while our phylogeny highlights that there is a general phylogenetic rationale behind the NPA classification, some considerable discrepancies are obvious. The Form group is clearly not monophyletic, something that is not unexpected given that this group is defined based on selection towards a specific body form, and the breeds included in this group have very distinct origins (e.g. heavy breeds originally developed for meat production and breeds originally selected for an improved homing performance). Despite its small number of breeds, the Wattle group is also not monophyletic, as the English Carrier and Dragoon cluster together with some Form breeds (e.g. Racing Homer). However, we do not find this surprising, since some of the ancestors of these Wattle breeds were used to create the modern Racing Homers (Tegetmeier 1871). While the Croppers & Pouters group form an almost monophyletic group they are rendered polyphyletic by the Marchenero Pouter. Even though previous studies have found Croppers & Pouters to be genetically allied (Stringham et al. 2012; Shapiro et al. 2013), Spanish Pouter breeds are morphologically distinct and inflate their crops differently from the other Croppers & Pouters. The Color group is also paraphyletic as it includes the Frillback breed, which is considered a Structure breed by the NPA. The Owls & Frills group is almost monophyletic, made paraphyletic by
the classification of the Chinese Owl as a Structure breed. Despite being the largest group, the TRHF group is mostly monophyletic, with the exception of the Mookee breed that is an outgroup to the Fantails (Structure), as previously demonstrated (Stringham et al. 2012; Shapiro et al. 2013). This pattern is not surprising, since the Mookee and Fantail breeds are closely related, and used to be known as the Narrow and Broad Tail Shaker, respectively (Moore 1735; Sell 2009). As also previously noted (Shapiro et al. 2013), the Trumpeter group is not monophyletic, possibly indicating that the Laugher derived voice is analogous to that found in Trumpeters, which would be in accordance with the diversity of this trait as different breeds in this group show different kinds of voices (e.g. drumming and laughing voices) (Marks 1975); however, further research is needed to test this hypothesis. Since the Structure group includes breeds with different genetic affinities, it is unsurprising that this group is made polyphyletic through inclusion of breeds that show close phylogenetic relationships with other groups, such as the Old Dutch Capuchine, which holds phylogenetic relationship with the Trumpeters. Finally, the Syrian group is paraphyletic as it includes the Egyptian Swift, which is considered a Form breed by the NPA.

Pigeon breeds are known for their variety of phenotypic traits, but it is not always clear when and how many times these traits emerged in their history of domestication (Supplementary Fig. 4). Based on both our and previous phylogenetic analyses (Stringham et al. 2012; Shapiro et al. 2013), we see that some traits seen in pigeon breeds are apomorphic, while others are spread across the entire phylogenetic space. As an apomorphic example, the ChestFrill trait is only found in the clade formed by Owls & Frills. On the other hand, the HeadCrest trait is found in breeds scattered across the phylogeny (e.g. Trumpeters and Structures), as is the FootFeathering trait (e.g. Croppers & Pouters, Trumpeters, Tumblers). The phylogenetic distribution of these traits is of paramount importance for future investigations attempting to reveal the genomic underpinnings behind the astonishing biological diversity that is present in pigeon breeds (Domyan and Shapiro 2017).

Genetic Variability of Pigeon Breeds

Domestic lineages have complex evolutionary histories shaped by population bottlenecks, strong artificial selection, reduced effective population sizes, and long periods of inbreeding that are occasionally punctuated by interbreeding among lines. As a consequence, these lineages generally have reduced levels of genetic diversity (Makino et al. 2018). To test whether pigeons also adhere to this trend, we took advantage of the fact that we have several samples for some of the breeds. Specifically, we calculated the observed levels of heterozygosity ($H_o$), nucleotide diversity ($\pi$), Watterson’s $\theta$ ($\theta_w$) and Tajima’s $D$ across the pigeon genome.

Values of $H_o$ were calculated for all samples in Dataset 1, except IndianFantail_03 and IranianTumbler_02, due to their inconsistent phylogenetic placement (see Supplementary Results & Discussion). All other statistics were calculated for the 12 breeds that had five or more individuals, as these genetic estimates only apply for population data (for the 23 triplicates only the WGS libraries were used). The individual $H_o$ levels among the pigeon breeds ranged from 0.0679% to 0.2395% (mean 0.1571%), with considerable variation within each breed and the presence of several outliers (Supplementary Fig. 6). In general, these values are similar with those reported for 7 duck breeds (mean 0.1530%), but are lower than those reported for two wild populations of mallard (mean 0.3009%) (Zhang et al. 2018), and are consistent with what would be expected of a lineage that
has been subject to the evolutionary forces imposed by the domestication process (Groeneveld et al. 2010; Makino et al. 2018). In comparing the triplicate samples, we noticed that each WGS-GBS sample invariably had the highest $H_o$, usually followed by the corresponding GBS library (Supplementary Spreadsheet). Even though we did not investigate this systematic difference in depth, we believe it relates to the previously observed excess of low frequency SNPs on GBS samples. Further investigation will be required to assess the significance of this issue, but we do not believe that it introduced any significant bias in our interpretation of the results. Moreover, although we were unable to identify any other obvious biases concerning the triplicates, we chose to be conservative and kept only the WGS libraries of each triplicate for all downstream analyses (ignoring both GBS and WGS-GBS).

The mean values of $\pi$ ranged from 0.0016 to 0.0027 (mean 0.0021), while the $\theta_w$ values ranged from 0.0013 to 0.0026 (mean 0.0019). Our estimates of $\pi$ are similar to those of other domesticated avian breeds (0.0020 to 0.0028) (Zhang et al. 2018), but considerably lower than those calculated for wild counterparts of domesticated avian species, such as the Mallard (approx. 0.0040) (Zhang et al. 2018) and the Red Junglefowl (0.0052) (Lawal et al. 2018), as would be expected considering the long history of extensive artificial selection experienced by the domesticated linages. Previously reported values for pigeon breeds (0.0036) (Shapiro et al. 2013) are higher than our estimates, but this might be due to the fact that the dataset included the resequenced genomes of both domestic and feral pigeons.

Next, to test for evidence of rapid population contraction (bottlenecks), we calculated Tajima’s $D$ for each breed. Estimates ranged from 0.1752 to 0.8428 (mean 0.4654), in accordance with reports for purebred lineages of other domestic animals such as quail (Wu et al. 2018) and sheep (Pan et al. 2018). These positive values probably reflect the recurrent history of bottlenecks inherent in the domestication process. Moreover, these values show a negative correlation with $H_o$ (Pearson correlation = -0.5128974; p-value = 0.08815), in agreement with lower variability during a bottleneck. The only exceptions are the English Carrier and Oriental Roller, which show relatively low values of both statistics. When these breeds are excluded, the correlation becomes stronger and highly significant (Pearson correlation = -0.859119; p-value = 0.001449). This could indicate that these breeds did not go through a very strong bottleneck during domestication but have been kept relatively isolated ever since. Nonetheless, this interpretation should be taken with care and further studies are required to fully understand these patterns. Interestingly, the Fantail, Indian Fantail and Chinese Owl breeds showed the highest Tajima’s $D$ and lowest genetic diversity according to $\theta_w$, suggesting that these breeds underwent a severe bottleneck. Unexpectedly, the Archangel and Starling breeds (both belonging to the Color group) showed the highest genetic diversity levels. Since their color patterns are the result of complex genetics, these breeds are not expected to be the products of outcrosses and, as such, would be relatively homogeneous. The Racing Homer also showed high genetic diversity, but this could be explained by i) its much larger effective population size given that it is raised in formidable numbers across the globe, and ii) it is a relatively young breed (approx. 200 years old) having being created out of many different breeds (Tegetmeier 1871).
Taken together, these results demonstrate that domestic pigeons adhere to the main trend of domestic lineages, showing reduced levels of genetic diversity probably originated from a recurrent history of population bottlenecks. Furthermore, likely due to variations in domestication periods, geographical origins, and domestication purposes, these results also indicate that artificial selection has imprinted different pigeon with different genomic signatures.

**Population Structure Across Pigeon Breeds**

Continued artificial selection on domestic lineages commonly leads to pronounced population structure, chiefly among established breeds and lines (e.g. ref. (Alves et al. 2015; Signer-Hasler et al. 2017)). Although previous studies have investigated the patterns of population structure among pigeon breeds, they examined either fewer genetic loci (32 microsatellites) (Stringham et al. 2012) or breeds (37 breeds) (Shapiro et al. 2013). As it has been demonstrated that genome-wide SNPs tend to better recapitulate evolutionary relationships in comparison with microsatellites (Väli et al. 2008; Gärke et al. 2012; Fischer et al. 2017), we used our larger dataset to unravel at a finer scale the patterns of population structure among pigeon breeds through the analyses of Proportions of Individual Ancestries (Admixture) and Multidimensional Scaling (MDS). Thus we created a new dataset including all samples except for the outgroup (C. rupestris, excluded due to its high divergence). This dataset consisted of 210 samples (Dataset 2; 161 GBS and 49 WGS) and yielded 26,082 SNPs with coverages ranging from 7,02X to 322,93X (mean 84.82X) and missing data from 0% to 1.89% (mean 0.24%) (Supplementary Spreadsheet).

Our Admixture (Fig. 2) and MDS (Fig. 3) analyses show considerable population structure among pigeon breeds, consistent with the findings of previous studies (Stringham et al. 2012; Shapiro et al. 2013). This likely arose as a direct product of continuous artificial selection. Interestingly, differences between the Fantails, Indian Fantails and Mindian Fantails and all remaining breeds are largely responsible for the first MDS dimensions and Admixture at K = 2, showing that these three breeds are the most divergent ones. It is worth noting that the Jacobin appears midway between Tumblers and Trumpeters on the MDS, further supporting its shared ancestry with both these groups. These analyses also provide extra evidence that some NPA groups seem to be relatively homogeneous and genetically isolated, such as the TRHF, the Owls & Frills, the Form, and the Wattle breeds. On the other hand, other groups are more genetically similar, such as the Croppers & Pouters and the Color breeds. The latter appears to include the English Trumpeter, but this breed is separated on the MDS dimension 3 (Supplementary Fig. 8). As also found by a previous study (Gazda et al. 2018), we highlight that a well-structured cluster is formed by the Racing Homers, despite the fact that we included samples from both Europe and North America. The Racing Homer breed was first established in Europe, and our results indicate that descendant populations in North America remained genetically similar. Furthermore, and also consistent with previous results (Stringham et al. 2012), both Feral samples had the greatest number of ancestry components at K = 20, as might be expected from an admixed feral population (Wang et al. 2017) (also ref. (Stringham et al. 2012) for a statement on this topic specifically with a reference to pigeons).
Inference of Admixture Events

The evolutionary history of pigeon breeds is rife with inter-breed crosses. Traditional population genetic analyses attempt to infer relationships among populations as a bifurcating phylogeny. However, simple bifurcating phylogenies may not correctly represent population histories (Cavalli-Sforza and Piazza 1975; Pickrell and Pritchard 2012). Thus, in an attempt to detect past admixture events among pigeon breeds, we employed a phylogenetic method that fits a population graph (allowing for both population splits and mixtures) to the allele frequency correlation patterns among a set of the sampled populations (Pickrell and Pritchard 2012). This analysis included all samples except the feral pigeon samples and the two samples deemed to be of uncertain provenance in the previous analysis (see Supplementary Results & Discussion), thus 207 samples (Dataset 3; 159 GBS and 48 WGS). After SNP calling, this dataset amounted to 26,504 SNPs with coverages ranging from 7.01X to 321.60X (mean 84.95X) and missing data from 0% to 1.94% (mean 0.24%) for genotype likelihoods and from 0.03% to 21.83% (mean 1.74%) for genotype calling (Supplementary Spreadsheet).

We found overall congruence between the ML (Fig. 1) and TreeMix phylogenies when no hybridization events were allowed (Supplementary Fig. 9a). When allowing for five hybridization events (Fig. 4), the first is from the Schmalkaldener Mohrenkopf to the node joining the Jacobin and the Old Dutch Capuchine (that now appears next to the Danish Tumbler, as previously seen (Shapiro et al. 2013)). The Jacobin is recognised as an ancient breed, and it was used to improve the feather length of the Schmalkaldener Mohrenkopf. Thus, we believe that the genomic affinity between these two breeds seen in our study as well as in a previous one (Stringham et al. 2012), might well explain this hybridization event. The second migration is from the Scandaroon to the node encompassing all Homers, the American Show Racer, the English Carrier and the Dragoon. The English Carrier is considered to be closely related to the Scandaroon and, given that both breeds share common ancestry with the breeds that were used in the creation of the Homers (Levi 1986), this is not unexpected (Levy 1965). The third migration is from the Syrian Dewlap to the node joining the Carneau and Scandaroon. Despite their lack of morphological similarity, these are thought to have originated (or have ancestors) in neighbouring regions in the Middle East (Moebes 1950). Thus, we believe that this migration could be due to a deep relationship relating to a common founder population. Both the fourth and fifth migrations are of note as they both recapitulate well documented facts. Specifically, the fourth migration is from the Polish Lynx to the node shared by all the Color pigeons. The Polish Lynx is known to be derived from a Field Pigeon (Color group) and a Cropper & Pouter pigeon (Schütte et al. 1971; Marks 1975), which could explain this migration. The fifth and final migration happens from the Marchenero Pouter to the California Color Pigeon. The latter breed was created very recently, and is yet to be recognised by the NPA. However, its creator (Frank Mosca) attests that the Marchenero Pouter was used in the creation of the California Color Pigeon (www.angelfire.com/ga3/pigeongenetics/ccpstandard.html). Given that the last two admixture events are well known and described, this provides an additional measure of confidence about the reliability of the other admixture events detected.
In summary, it is clear from our results that the major admixture events underlying the creation of pigeon breeds can be reconstructed with genomic data, and thus as more datasets are generated, much more will be learnt about the history of many other breeds.

**Linkage Disequilibrium Levels Across Pigeon Breeds**

Patterns of Linkage Disequilibrium (LD) decay can greatly affect the efficiency of Genome-Wide Association Studies (GWAS). While this has been previously explored using a reduced number of samples and breeds (Shapiro et al. 2013), we took advantage of our expanded dataset to improve estimates of pigeon LD decay, something that may help clarify the appropriateness of different pigeon datasets in studies attempting to link genetic variation with phenotypic variability. Specifically we used all samples in Dataset 3 (only excluding the outgroup) to calculate $r^2$, but focused on the 0.95 quantile since these data points are most relevant for GWAS. Our analyses yield higher LD estimates among pigeon breeds than those previously reported (Shapiro et al. 2013) (Supplementary Fig. 10), with median relevant LD (Aerts et al. 2007) decaying at around 30 kb, but possibly extending until 200 kb due to high variance.

Although the average inter-SNP distance is approximately 17 kb (Supplementary Fig. 7b), we conclude that the levels of long range LD found here indicate that our marker density is appropriate for GWAS based on GBS data.

**Assessing GBS Performance for Genome-wide Association Studies**

Despite the sparse set of markers that are intrinsically generated by reduced-library sequencing protocols, several studies have demonstrated that this kind of data can be used to successfully perform GWAS (e.g. ref. (Parker et al. 2016; Otto et al. 2017; Barria et al. 2018)). In this way, we wished to investigate whether GWAS analyses on our GBS-derived panel of SNPs would be able to recapitulate the results of previous analyses undertaken on WGS data. Thus, we undertook a GWAS analysis using Dataset 3 targeting two morphological traits for which the genomic basis was already reconstructed using WGS. A previous study (Domyan et al. 2016) described two haplotypes (that include regions: Scaffold_79:6719000-6763000 in Cliv_1.0, ScoHet5_1033.2:7486712-7530712 in Cliv_2.1; and Scaffold_70:7313000 in Cliv_1.0, ScoHet5_149:1203848 in Cliv_2.1) to be linked to the FootFeathering trait. In our dataset we found three SNPs on scaffold ScoHet5_1033.2 and one SNP on scaffold ScoHet5_149 to be significantly associated with this trait (Fig. 5). Remarkably, the three SNPs on scaffold ScoHet5_1033.2 were at approximately 135 kb from the known haplotype, while the single SNP on scaffold ScoHet5_149 was at approximately 300 kb from the other known haplotype. Thus, considering the intrinsic scale of each method, our GBS-based GWAS was able to identify the same genomic regions involved in the FootFeathering trait that were previously reported. Intriguingly, we also found a SNP (ScoHet5_205:16892632) on a third scaffold significantly associated with this trait at a marginal level. We speculate that it represents either i) a false positive, ii) a scaffold physically close to one of the two known haplotypes associated with this trait, iii) a third but yet unidentified locus that contributes to the trait. As for the HeadCrest trait, an 11-kb haplotype containing a putatively causative SNV (Scaffold_612:596613 in Cliv_1.0;
ScoHet5_3280:262233 in Cliv_2.1) is thought to be associated with this trait (Shapiro et al. 2013). Intriguingly, no SNP in our dataset was significantly associated with this trait (Supplementary Fig. 11). However, there is a markedly low density of GBS cut-sites (and hence SNPs) around the region in question (only 3 SNPs less than 300 kb away), leading to a considerably greater SNP spacing in relation to the haplotype we are trying to detect.

Our results corroborate a previous studies that demonstrate the feasibility of successfully performing GWAS based on reduced-representation library data (e.g. ref. (Parker et al. 2016; Otto et al. 2017; Barría et al. 2018)). Nonetheless, given the wide SNP spacing produced by this sort of method (e.g. GBS), it is important to highlight that the power of this type of analyses will be strongly dependent on the average distance between SNPs, as well as levels of LD. In this regard, protocols with lower levels of missing-data (e.g. paired-end and double-digestion RADseq) are expected to perform considerably better.
Conclusion

As a result of their rich and complex history of continuous artificial selection, the NPA (www.npausa.com) today recognises approximately 230 pigeon breeds whose phenotypes exhibit incredible diversity. With the goal to further promote GWAS and comparative genomics studies on pigeon breeds, we conducted the most inclusive genomic study to date for the group. In doing so, we demonstrate a framework for analysing GBS and WGS data under a common framework without the introduction of any significant bias. To the best of our knowledge, this is the first study that successfully analysed these two kinds of genomic data following a unified approach.

Our results demonstrate that there is considerable population structure across pigeon breeds as a result of intensive artificial selection. In addition, we have demonstrated that pigeon breeds can indeed be classified into distinct groups with different levels of genetic homogeneity and evolutionary histories. Furthermore, our results corroborate previous studies which showed that while some derived traits present in pigeon breeds were probably inherited from a common ancestral breed, others are distributed across the phylogeny (probably due to intentional transfer of traits from one breed to another) (Shapiro et al. 2013; Domyan et al. 2014; Domyan et al. 2016; Vickrey et al. 2018). Nonetheless, since the majority of the pigeons analysed in our study were collected outside the regions where their respective breeds were created, we caution that our results should be interpreted with care since these breeds might have been considerably altered once exported from their place of origin (Parker et al. 2017). Thus, we advocate that future genomic studies on breeds should strive to sample individuals within the breeds’ respective regions of origin. Despite this potential methodological caveat, we believe that our study is an important step towards the elevation of the domestic pigeon as a model organism for genomic investigations (Domyan and Shapiro 2017).

In this way, we envisage that the generation of full genomes for all recognised breeds would be a milestone towards this goal, and that such achievement would benefit investigations in several branches of the biological sciences where the pigeon has been used as a model system. More specifically, it would undoubtedly facilitate the proliferation of GWAS and comparative genomics studies taking advantage of the entire assortment of biological features seen in the group, as has been performed for other domestic animals (Axelsson et al. 2013; Imsland et al. 2016; Carneiro et al. 2017; Alberto et al. 2018). Furthermore, we believe that the pigeons short generation time, easy animal handling and relatively small genome compared to other model organisms, places this group at the privileged position for scientific queries as was foreseen by Charles Darwin more than 150 years ago.
Methods

Classification of Pigeon Breeds

Although many countries have their own national pigeon associations that classify pigeon breed into specific groups, we elected to follow the classification from the National Pigeon Association of the United States of America (NPA; www.npauusa.com) given its status as one of the largest associations of pigeon breeders. Additionally, the NPA publishes a book that describes and classifies all of its recognized pigeon breeds. According to NPA’s 2010 Book of Standards, all recognized breeds are divided into nine groups, exclusively depending on their function, morphology, vocal abilities, origin, etc. These groups are named Form (includes mostly homing related breeds being mainly selected for their body form), Wattle (relatively small group including breeds having pronounced wattles), Croppers & Pouters (breeds that present conspicuous crops), Color (breeds chiefly developed for their colors and markings), Owls & Frills (breeds showing a chest frill), Trumpeters (a group of breeds selected for voice characteristics), Tumbler, Rollers & High Flyers (TRHF; the largest group covering all breeds with a performing background), Structure (diverse group of breeds being selected for their exuberant ornamentation) and Syrian (these breeds are grouped together mainly due to their common geographical origin).

In order to encompass as much as the pigeon diversity as possible, we selected 55 recognized pigeon breeds from all nine NPA groups (based on their worldwide popularity, morphological variety, and breed history), as well as 6 other breeds not currently recognized by the NPA.

Genomic Data

This study is based upon the merging and analysis of two different kinds of genomic data. The first derives from three published datasets of whole-genome sequences (WGS), while the second is a newly generated GBS data. The raw WGS data consisted of 39 purebred pigeons, 2 feral pigeons and 1 out-group (Columbia rupestris, Pallas 1811) (Shapiro et al. 2013), 2 Pomeranian Pouters (Domyan et al. 2016) and 8 Racing Homers(Gazda et al. 2018) (Supplementary Spreadsheet) as downloaded from a public database (NCBI; Project Numbers: PRJNA167554, PRJNA284526 and PRJNA427400, respectively). Each SRA file was converted into FASTQ format using fastq-dump from SRA Toolkit v2.7; https://github.com/ncbi/sra-tools, using default parameters plus options --split-files and --skip-technical.

Reduced-Representation Library Sequencing Data

Genotyping-by-Sequencing (GBS)

We generated GBS data for 190 samples representing 61 breeds from the collection of the Shapiro Lab at the University of Utah. Genomic DNA extractions were performed using the DNeasy Blood & Tissue Kit® (Qiagen, Valencia, CA) following manufacturer’s instructions. The extracts were quantified using the Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA) with no modification of the standard protocol. To check for molecular integrity, each DNA extract was run on a 1% agarose gel against a 1-kb ladder. Samples were sent to...
the Institute for Genomics Diversity – Cornell University, where the GBS method was performed following the original protocol (Elshire et al. 2011). We submitted 190 extracts that passed our filters (minimum DNA concentration of 10 ng/μL and average fragment size above 20 kb), split into two 96-well plates (PBGB_1 and PBGB_2). A negative control (water) was included in each plate in a predetermined well. At Cornell, the DNA samples were treated with the restriction enzyme EcoT22I before library preparation. Then, the quality of each library was inspected through the visualization of their fragment size distributions. All libraries passed quality control (appropriate concentration, fragment size distribution and minimal adapter dimers). The respective libraries of each plate were pooled separately and then sequenced on two runs of the HiSeq 2000 apparatus (Illumina, San Diego) under a protocol of single-end reads of 100 bp.

**GBS Data Demultiplexing**

We used the software GBSX v1.3 (Herten et al. 2015) to demultiplex our GBS data allowing for 1 mismatch in the barcodes (-mb 1), 1 mismatch in the enzyme cut-site (-me 1), and ensuring that no common sequencing adapter was to be removed (-ca false).

**Read Filtering and Mapping**

The software Paleomix v1.2.5 (Schubert et al. 2014) was used to filter and map our entire dataset. The two datasets (both WGS and GBS) were run with the exact same parameters (see below for details on parameters used for specific programs), except for the removal of PCR duplicates (that was only performed for WGS samples), and the sequencing adapters (since GBS did not use the common Illumina adapters). For the reference genome, we used a preliminary version of the Cliv_2.1 pigeon assembly (Holt et al. 2018) [https://sid.erda.dk/wsgi-bin/ls.py?share_id=ArXpW64Hxt].

**Filtering for GBS Chimeric Reads**

During our initial inspection of the data, we noticed that some GBS reads seemed to be chimeric. Specifically, the merging of reads derived from two or more biological cut-sites into one single artificial read (Supplementary Fig. 1). We did not fully investigate these abnormal cases in the current study, but we suspect that this technical issue is caused by the undesired ligation of some cut-sites to other cut-sites during the adapter ligation step. In order to be conservative, we excluded all chimeric reads as they could bias our coverage statistics. Briefly, these were defined as those reads with i) more than one cut-site, ii) mapped to two or more non-contiguous regions in the genome.

**Analysed Genomic Fraction**

We restricted our analyses to only the fraction of the genome theoretically available to the GBS method. To determine this fraction, we performed an in-silico digestion on the Cliv_2.1 reference assembly with the same
enzyme used in our GBS protocol (EcoT22I) by employing BioSeq v1.11 (Cock et al. 2009), and considered only the regions spanning 92 bp downstream and upstream each locus. Importantly, since some loci were located less than 92 bp apart from each other, we merged these specific loci into single locus. Hereafter, our final set of loci will be referred to as Merged_Loci.

### Trimming of Reads

AdapterRemoval v2.1.7 (Schubert et al. 2016) was used to filter low quality reads, trim low quality read fragments, and remove adapters using default parameters, except for: a minimum read length of 30 bp (--minlength 30), collapse paired-end reads (--collapse yes), remove stretches of Ns (--trimns yes), remove consecutive stretches of bases with qualities below 15 (--trimqualities yes, --minquality 15), and discard reads with more than 40 Ns after trimming (--maxns 40).

### Mapping

The software BWA v0.7.15 (Li and Durbin 2009) was used to map the reads against the Cliv_2.1 reference assembly using the algorithm BWA-MEM, ignoring all reads with mapping quality below 20. Finally, to minimize increased error rates around indels, we used the software GATK v3.6 (McKenna et al. 2010) to perform indel realignment. We used PaleoMix to generate mapping statistics for all loci the set Merged_Loci. Moreover, in order to explore if GBS data could be merged with WGS data without the introduction of systematic biases, we created a third replicate for each of the 23 replicated samples by merging the each WGS and GBS BAM files; these joint samples are hereafter referred as WGS-GBS replicates.

### Molecular Sexing

Since the Cliv_2.1 reference assembly is only at the scaffold level, we took advantage of a chromosome level pigeon assembly (Damas et al. 2017) in order to identify which scaffolds belong to chromosomes 6 and Z. To do so, we independently blasted these two chromosomes against our reference genome using Blast+ v2.6.0 (Camacho et al. 2009), requiring a minimum percentage identity (-perc_identity 95), e-value (-evalue 1e-5), alignment length (greater than 1 kb), subject length (greater than 10 kb), and a proportion of gaps in the alignment smaller than 1%. All subject sequence with alignments that passed the filters were assumed to belong to the respective query chromosome (either 6 or 2), and their coverage were calculated.

### Data Filtering

#### Filtering of Failed Samples

In order to filter out those samples for which a minimum number of reads was not produced, we generated a presence/absence matrix for all the loci comprised in the set of loci Merged_Loci (presence if a locus was covered by 3 or more reads). Due to the magnitude of the matrix, we clustered the loci (k-means with K = 300
clusters), and plotted the matrix as a heatmap with the samples hierarchically clustered by employing the R package `pheatmap` (Kolde 2012). We then inspected this heatmap by eye and decided to remove from further analyses the samples in the entire tip branch where the GBS negative controls were present.

**Filtering of Possible Paralog Loci & Genome Portion Analysed**

We took advantage of the WGS dataset to flag possible paralog loci through the conventional methods of loci exclusion based on an excess of Global Depth (GD) relative to the mean. Briefly, we first used the software `ANGSD v0.921` (Korneliussen et al. 2014) to calculate the GD per base pair of each GBS locus for all the WGS samples and, considering that the GD distribution follows a Poisson distribution, excluded those loci with GD considerably higher than average (> 800X) (Supplementary Fig. 3a). If not stated otherwise, this and all following plots were created using the R package `ggplot2` v2.2.1.9 (Wickham).

**Data Analysis**

We generated specific datasets to serve as inputs of the analyses conducted by performing multiple runs of data analysis using of the package `ANGSD v0.921` (Korneliussen et al. 2014). Although each of these runs had their own specificities, all of them obeyed to some general parameters and conditions. First, only the set of loci Merged_Loci, and scaffolds longer than 1 kb (4,063 scaffolds) were analysed, in order to avoid analysing regions of problematic assembly (e.g. repetitive regions). Second, several filters were applied for minimum mapping quality (`-minMapQ 30`), minimum base quality (`-minQ 20`), missing data (`-minInd 95%`), GD (`-setMaxDepth 275X per individual`), minimum genotype posterior probability (`-postCutoff 0.95`), minimum minor allele frequency (`-MinMaf 0.005`), remove anomalous reads (`-remove_bads 1; SAM flag above 255`), adjust mapping quality for excessive mismatches (`-C 50`), perform BAQ computation (`-baq 1`), minimum coverage for genotype calling (`-geno_minDepth 3`), use `SAMtools` genotype likelihood model (`-GL 1`), and estimate posterior genotype probabilities assuming an uniform prior (`-doPost 2`). For runs where SNP calling was performed, we used the `ANGSD` SNP calling method (`-SNP_pval 1e-6`), where a Likelihood Ratio Test is used to compare between the null (maf = 0) and alternative (estimated maf) hypotheses by using a chi-square distribution with one degree of freedom.

**Genetic Diversity**

We followed the instructions provided by `ANGSD v0.921` (Korneliussen et al. 2014) to calculate the unfolded global estimate of the Site Frequency Spectrum (SFS) in order to calculate the observed fraction of heterozygous sites ($H_o$) per sample, as well as the estimates of nucleotide diversity ($\pi$), Watterson’s $\Theta$ ($\Theta_w$) and Tajima’s D (per breed) (Korneliussen et al. 2013). The observed fraction of heterozygous sites was calculated as the ratio between the number of heterozygotes and the total number of sites with information in percentage.
**Phylogenetic Reconstruction**

For the Maximum-likelihood (ML) phylogenetic reconstruction, we used RAxML-NG v0.5.1b (https://github.com/amkocz/raxml-ng) to perform two phylogenetic searches using as starting topology either a NJ phylogeny or 20 random topologies. The Neighbour-joining (NJ) phylogenetic reconstruction was based on a pairwise genetic distances matrix calculated directly from the genotype likelihoods outputted by ANGSD using the software ngsDist v1.0.2 (Vieira et al. 2016) with pairwise deletion (--pairwise_del), and inferred using the software FastME v2.1.5 (Lefort et al. 2015) with the SPR tree topology improvement (-s). Both these searches employed the GTR model with discrete GAMMA with 4 categories, mean category rates and ML estimate of alpha (--model GTR+G), as well as used the site repeats optimization option (--site-repeats on). We chose the phylogeny with the highest likelihood (that incidentally turned out to be the one starting from the NJ phylogeny) and used RAxML-NG to calculate bootstrap values using the bootstrap option based on 100 replicates (--bs-trees 100) and the same setup model used to compute the main phylogeny. The final bootstrapped phylogeny was visualised and plotted using the online software iTOL v4.0.3 (Letunic and Bork 2016).

**Multidimensional Scaling**

We calculated a pairwise genetic distances matrix in the same aforementioned way and used it to conduct a Multidimensional Scaling (MDS) analyses using the R package cmdscale.

**Estimation of Individual Ancestries**

The software ngsAdmix v32 was used to estimate proportions of individual ancestries for K = 2 up to K = 20 in 100 replicates using default parameters, except for tolerance for convergence (-tol 1e-6), log likelihood difference in 50 iterations (-tolLike50 1e-3), and maximum number of EM iterations (-maxiter 10000).

**Inference of Migration Events**

We ran TreeMix v1.13 (Pickrell and Pritchard 2012) using default parameters, except for size of block for estimation of covariance matrix (-k 100), sample size correction (-noss), round of global rearrangements after adding all populations (-global), and setting the Crupestris_01-WGS samples as the out-group (-root Crupestris_01-WGS). Migration edges were added until residuals did not appreciably decrease (five in our case). The results were plotted using the R function plotting_funcs provided by TreeMix.

**Linkage Disequilibrium**

We used the software ngsLD (https://github.com/fgvieira/ngsLD) to estimate the decay of linkage disequilibrium (LD) based on genotype likelihoods using sites with 0.01 as minor allele frequency (--min_maf 0.01). We plotted the r² estimates using the fit_LDdecay.R script provided by ngsLD setting 500 kb as maximum distance between SNPs (--max_kb_dist 500), exhaustive fitting (--fit_level 100) and a fitting bin size of 200 bp (--fit_bin_size 200).
**Genome-Wide Association Study**

From the genotype likelihoods, we calculated the expected genotypes and used these as allele dosages. Then, we employed the software GEMMA-v0.96 (Zhou and Stephens 2012) to calculate a centered relatedness matrix (-gk 1) excluding SNPs with minor allele frequency below 0.01 (-maf 0.01). Based on ref. (Levi 1996) and the NPA 2010 Book of Standards, we created a phenotype table scoring both the HeadCrest and the FootFeathering traits across the different breeds (Supplementary Spreadsheet). Finally, we used the genetic relatedness matrix to fit a linear mixed model in GEMMA, and performed a likelihood ratio test concerning each of the scored traits separately. Moreover, for both association analyses conducted, we performed a permutation test using MVNpermute (Abney 2015) in order to determine an appropriate significance threshold for each of the association analyses by running 100 permutation replicates and then independently running GEMMA on each permuted phenotype matrix. The p-values from all 100 GEMMA runs were concatenated, and then the 5th percentile p-value, corrected for the total number of tests, was used to draw the lines of significant of each study. Results were plotted using the R package ggman v0.99.0.beta (https://github.com/drveera/ggman).
Funding

This project was funded by the Lundbeck Foundation (Award: R52-5062) and the European Research Council (Consolidator Grant: 681396) granted to MTPG, a Ciência Sem Fronteiras Full PhD Abroad Scholarship (Grant: 201761/2014-9) granted by the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) to GP, and grants from the National Institutes of Health (R01GM115996) and National Science Foundation (CAREER DEB1149160) to MDS.

Competing Interests

The authors declare no competing interests.

Acknowledgments

We thank the Cornell University Biotechnology Resource Center for its genotyping services, especially Sharon E. Mitchell and all lab technicians involved; current and former members of the Shapiro lab for collection of samples used in this study; and members of the National Pigeon Association and Utah Pigeon Club for advice and permission to collect samples from their birds. Samples were collected in accordance with protocols approved by the University of Utah Institutional Animal Care and Use Committee (protocols 10–05007, 13–04012, and 16–03010). We also thank the University of Washington Burke Museum for the *C. rupestris* tissue sample (UWBM 59803, the same sample previously reported in Shapiro et al. 2013).

Author Contributions

G.P., F.G.V., M.D.S. and M.T.P.G. conceived the study. M.D.S and M.T.P.G. obtained chief financial support. M.D.S. and his lab group collected, curated, and provided all samples. G.P. performed the DNA extractions and quality checks. G.P. and F.G.V. conducted the computational analyses. G.P., F.G.V., M.T.P.G. and H.vanG. interpreted the results assisted by M.D.S. The first draft of the manuscript was written by G.P. and F.G.V. with input from M.D.S., H.vanG. and M.T.P.G. All authors critically reviewed and approved the final manuscript.

Data Availability

All demultiplexed GBS sequencing data is publicly available at NCBI (SRA Project Number: PRJNA485426), while the raw data is stored on the University of Copenhagen’s long term storage server (ERDA) and can be downloaded from https://sid.erda.dk/wsgi-bin/is.py?share_id=BTZBGV6Tsj.
References


Darwin C. 1855. Letter no. 1772. Darwin Correspondence Project [Internet]. Available from: http://www.darwinproject.ac.uk/DCP-LETT-1772

Darwin C. 1968. The Variation of Animals and Plants under Domestication. London, John Murray


Figures Legends

Fig. 1 | Maximum-likelihood phylogeny of pigeon breeds. Phylogeny describing the relationships between over 200 pigeon individuals representing 67 breeds and two feral pigeons. The outgroup (C. rupestris) is depicted in red, while the 23 triplicates are highlighted in purple. The colored ring depicts the NPA group of each sample, whilst the outermost circles and stars represent the traits present in each breed (information is given just for one sample of each triplicate). Nodes with bootstrap values above 70% are marked with green circles. The inner circular lines represent the inverted scale.

Fig. 2 | Admixture proportions of pigeon breeds. Individual are represented by columns, while rows depict the Admixture proportions based on the assumption of different numbers of ancestral populations (K = 2 - 20). Individuals are sorted by breeds (gray upper labels) and grouped per NPA groups (coloured upper labels; the colours used are as in the phylogeny presented in Fig. 1).

Fig. 3 | Multidimensional Scaling analysis of pigeon breeds. Dimensions 1 and 2 are plotted and each point on the plot represent a single individual. The colored solid ellipses represent the rough distribution of the most homogeneous NPA groups (the colours used are as in the phylogeny and Admixture plots), while the dashed ellipse depict the distribution of the Fantail breeds.

Fig. 4 | TreeMix maximum-likelihood phylogeny of pigeon breeds. Five migration events among different pigeon breeds are represented by arrows on the phylogenetic graph. The scale bar indicates ten times the average s.e. The out-group is marked in red. The model residuals are plotted in Supplementary Fig. 9.

Fig 5. | Manhattan plot for the GWAS analysis for the FootFeathering trait in pigeon breeds. The green line indicates the permutation derived significance threshold for association (P < 1.91 X 10^-5). The genomic coordinates of SNPs with significant associations to the phenotype are specified. The Y-axis shows the ~log10-transformed two-tailed P-value of each SNP from the GWAS meta-analysis (of linear and logistic regression statistics), while the X-axis shows indexed positions along the scaffolds.
### Molecular Biology and Evolution

#### Fig. 2

Fig. 3

**Dimension 1 (8.54%)**

- American Giant Homer
- American Show Racer
- Cameo
- Carneau
- Egyptian Swift
- King
- Laker
- Maltaese
- Polish Lynx
- Racing Homer
- Rust
- Show Type Homer

**Dimension 2 (4.77%)**

- American Flying Tumbler
- Berlin Long-faced Tumbler
- Budapest Tumbler
- Castilian Tumbler
- Colombine
- Dorking
- English Long-faced Tumbler
- Medium-faceted Crested Helmet

**Form**

- Croppers & Pouters
- Tumblers, Rollers & High Flyers

**Color**

- English Pouter
- Croppers & Pouters
- Structure

**Wattle**

- Barb
- Dragon
- English Carrier
- Scandinavian
- Spanish Barb

**Owls & Frills**

- American Flying Tumbler
- Ancient Tumbler
- Indian Fantail

**Non-NPA Breeds & Ferals**

- Backs Tumbler
- Bermuda Roller
- California Color Pigeon

**Structure**

- Chinese Owl
- Fantail
- Frels
- Syrian Fantail
- Syrian Develop
Darwin’s Fancy Revised: An Improved Understanding of the Genomic Constitution of Pigeon Breeds

George Pacheco1*, Hein van Grouw2, Michael D. Shapiro3, M. Thomas P. Gilbert1,4 & Filipe Garrett Vieira1

1Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark.
2Bird Group, Department of Life Sciences, Natural History Museum, Akeman Street, Tring, Herts. HP23 6AP, United Kingdom.
3School of Biological Sciences, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112, USA.
4NTNU University Museum, Norwegian University of Science and Technology, NO-7491, Trondheim, Norway.
*george.pacheco@snm.ku.dk

Table of Contents

Supplementary Methods ........................................................................................................... 2
Supplementary Results & Discussion ........................................................................................ 2
Sequencing Output & Data Filtering ........................................................................................ 2
Molecular Sexing Based on GBS Data ....................................................................................... 3
Phylogenetic Description ........................................................................................................... 3
Within/Between Genetic Distances of Pigeon Breeds .............................................................. 3
Abnormal Samples ................................................................................................................... 4
Notes on Breeds ....................................................................................................................... 4
Notes on the NPA Classification ............................................................................................. 4

Supplementary Figures ............................................................................................................ 5
Supplementary Fig. 1 | Schematic example of a GBS chimeric read. ........................................ 5
Supplementary Fig. 2 | Heatmap based on the presence/absence matrix. ............................... 6
Supplementary Fig. 3 | Global Depth (GD) density plots. ......................................................... 7
Supplementary Fig. 4 | Examples of phenotypic diversity among pigeon breeds. .................... 8
Supplementary Fig. 5 | Violin plot of genetic distances. ............................................................... 9
Supplementary Fig. 6 | Proportion of observed heterozygous sites. ............................................ 10
Supplementary Fig. 7 | Sites and SNPs information. ................................................................. 11
Supplementary Fig. 8 | Multidimensional Scaling Analysis (MDS). ............................................ 12
Supplementary Fig. 9 | TreeMix phylogenies for pigeon breeds and model residuals. ............. 13
Supplementary Fig. 10 | Linkage-disequilibrium (LD) across pigeon breeds. ......................... 19
Supplementary Fig. 11 | Manhattan plot for the GWAS analysis for the HeadCrest trait. ........ 20

Other Supplementary File
PBGP Supplementary Spreadsheet
Supplementary Methods

Since the first lane of sequencing on plate PBGB_1 did not produce the minimum desired number of reads (250 M), this whole plate was re-sequenced. Even after this extra sequencing on plate PBGB_1, columns 1-3 of this plate as well as of plate PBGB_2 did not produce enough reads. Thus, the samples placed in these wells were pooled together and re-sequenced once again on a separate lane. The reads of samples having reads coming from different lanes were merged together using the Lane option of the PaleoMix software (Supplementary Spreadsheet).

Although we caution that bugs might emerge due to changes in software versions, the entire pipeline for this project alongside with its documentation to serve as a guideline for the analyses herein performed can be found at https://github.com/layka-pacheco.

Supplementary Results & Discussion

Sequencing Output & Data Filtering

We generated 843,216,251 demultiplexed GBS raw reads for 190 samples, representing 61 breeds. The percentage of chimeric reads removed ranged from 0% to 2.65% (average of 0.87%) per sample. The calculated coverage across the regions of interest ranged from 0X to 13.83X (average of 6.12X) among the GBS samples, while it varied from 1.14X to 64.62X (average of 13.20X) among the 52 WGS samples (Supplementary Spreadsheet).

Based on a presence/absence heatmap (Supplementary Fig. 2), we excluded 8 samples from downstream analyses (6 GBS and 2 WGS) due to inefficient sequencing. This heatmap shows that there was large discrepancy in the levels of sequencing effectiveness across all predicted cut-sites. An explanation for this could be the large distances between GBS loci that would lead to an indirect size selection at PCR and sequencing steps. In fact, the average distance between every predicted cut-site and the next was 2,785.51 bp. Thus, assuming that fragments longer than 500 bp were excluded during the GBS experimental procedures due to PCR restrictions (extension step lasted 30 sec and the Taq polymerase used is expected to extend 1 kb/min), we expected that roughly 80% of all GBS loci were actually inaccessible as 80.54% of the cut-sites produced fragments greater than 500 bp. This number is very close to our observed missing data value of 74.57% (loci with no data for any sample). Therefore, even though we cannot rule out the role of additional factors behind this observation, we warn that results of in-silico digestion performed as part of a GBS experimental design should be considered with caution as the real sequencing efficiency may be rather different from what was predicted.

Sequence variation between paralogous genomic regions can lead to the identification of false-positive SNPs. Thus, standard genotyping pipelines usually attempt to remove these regions from downstream analyses. The most commonly used method is to remove all regions with extremely high coverage relative to the mean. However, since it is not possible to identify PCR duplicated reads in single-end GBS data, the Global Depth (GD) distribution can be rather misleading. Other methods have been developed to infer paralogous regions in GBS and similar datasets, but they assume a natural population (e.g.69). Since pigeon breeds are not natural populations, we took advantage of the presence of both WGS and GBS samples in our data to develop a new approach to identify paralogs (see methods), resulting in the exclusion of 0.46% of the loci. In order to validate this approach using the whole dataset, we plotted the average GD for all loci on top of the average GD just for the loci flagged as possible paralogs by our approach (Supplementary Fig. 3b). Intriguingly, the possible paralogous loci identified through the WGS dataset showed no correspondingly high GD on the entire dataset. We believe this is a result of the aforementioned extreme difference of sequencing performance across the cut-sites. Therefore, we conclude that excluding possible paralog loci simply based on higher than expected GD is not efficient and more robust methods69 should be used whenever feasible. Post-filtering, we retained 354,919 loci (covering roughly 5.87% of the pigeon genome) which were used in all downstream analyses.
Molecular Sexing Based on GBS Data

In order to check if the GBS data would be appropriate to perform molecular sexing, we compared the coverage between the sexual chromosome Z (haploid in female birds) and an autosome of similar size (chromosome 6) for all our samples, under the assumption that approximately half the number of reads would map to the Z chromosome than to the similarly sized autosome in female birds. The calculated values ranged from 0.498 to 1.085 with very few numbers being too far away either from 0.5 or 1.0, thus we could confidently score all samples as either male or female. While 28.2% of the samples scored as female, 71.8% scored as male. Moreover, we noticed some incongruences between our GBS sexing and the phenotype sex noted at the time of sampling, which is consistent with the notorious challenge of sexing most pigeon breeds solely based on morphological and behavioural traits, especially juveniles or young adults. Out of 173 cases for which we had both score types (only samples that passed to the variant calling phase), 14.45% of the cases were incongruent (Supplementary Spreadsheet).

Phylogenetic Description

The outermost clade of our phylogeny encompasses all Color breeds plus the Frillbacks (Structure), followed by two sister clades with the Jacobin, Old Dutch Capuchine, and Schmalkaldener Mohrenkopf (all Structure), and most Trumpeters. A third clade is formed by the Maltese, King, Polish Lynx and Runt (all Form), plus all Croppers & Pouters, and the Laugher (Trumpeter). The Fantails (Structure), the Shakhsharli (Syrian), the Lahore (Form), the Mookee, and one Iranian Tumbler (both TRHF) form the forth clade, followed by the fifth clade formed by most TRHF, the Cumulet, the Temeschburg Schecken (also known as Timisoara Tumbler), and the Medium-faced Crested Helmet (all part of TRHF). The sixth main clade is formed by the Homers (Form), Carriers, Dragoon (both Wattle), and the Ferals, supporting previous claims that current American feral pigeon populations could have been originated mostly from stray Homers11,12. The Barb, the Spanish Barb and the Scandaroon (all Wattle breeds), together with the Carneau (Form) form a new (i.e. not reported before) sister clade to the sixth clade. The final and seventh clade seen in our phylogeny is formed by the Chinese Owl (Structure), the rest of Owls and the Oriental Frill (both part of Owls & Frills), the Lebanon and Syrian Dewlap (both Syrian), and the Egyptian Swift (Form).

Within/Between Genetic Distances of Pigeon Breeds

In order to better identify putative phylogenetic outliers of each breed, we summarized all the pairwise genetic distances on a violin plot (Supplementary Fig. 5), clustering the data into 4 categories: Intra-replicates (distances among WGS, GBS and WGS-GBS triplicates), Intra-breeds (distances among individuals within each breed), Inter-breeds (distances among individuals belonging to different breeds) and Inter-species (distances from all pigeon samples to the outgroup). As seen in the phylogeny, samples IndianFantail_03 and IranianTumbler_01 fall within the Cumulet and the Shakhsharli clades, respectively. The pattern can be seen even more clearly in the violin plot, where these samples show very high intra-breed distances (at an inter-breed level), and low inter-breed distances to the Cumulet and Shakhsharli, respectively (at an intra-breed level). Interestingly, despite the fact that IndianFantail_03 sample was deemed an Indian Fantail by its owner, the breeder also reported that this pigeon had substantial Cumulet ancestry. On the other hand, given their shared geographic region of origin, the close genetic similarity between the IranianTumbler_01 and the Shakhsharli breed is completely expected.

Another unexpected result is the very low intra-breed genetic distances (at a replicate level) between Fantails 02 and 09, and between Scandaroons 01 and 02. These results could indicate that these samples either derive from the same individuals, identical twins, or siblings of highly related parents (breeders often breed for several years from a pair that produces good offspring). Since these four pigeons had different leg-band numbers (a unique tag for each bird that is not changeable or re-
movable) and avian identical twins are extremely rare, we postulate that these samples might in fact represent siblings from closely related parents (e.g. the Fantails belonged to the same breeder). Finally, there are three pairs of breeds that show very high genetic similarity (at an intra-breed level): Jacobins and Old Dutch Capuchines, English Long-faced Tumbler and Parlor Roller, and Laugher and Marchenero Pouter. The two breeds in the first pair are believed to share a common origin and are also morphologically similar (e.g. both present a well-developed hood), thus it is not surprising that they are genetically so similar. The second pair, both belonging to the TRHF group, thus we speculate this could be the result of shared ancestry or recent inter-breed cross. In fact, it has been reported that dog breeds outside their country of origin i) experience a loss in genetic diversity due to importation bottleneck, and ii) experience more inter-breed admixture, since breeders in the importing countries tend to be more prone to perform experimental crosses. This could well be similar with pigeon breeds. Thus, since most of our samples were collected in the USA, which is not the native country for the majority of our breeds, the pigeons within our panel of samples might be more outbred than pigeons belonging to the same breeds in their respective country of origin.

**Notes on Breeds**

Breeders have outcrossed the Indian Fantail with other breeds (namely small Tumblers) in order to achieve the Mindian Fantail miniaturised standard (personal communication from D. Skiles to M. D. S.); the Jacobin was used to improve the feather length of the Schmalkalender Mohrenkopf (personal communication from breeders to H. van G.).

**Notes on the NPA Classification**

Based on the results of the phylogenetic, Admixture and MDS analyses, we are in the position to highlight some points regarding the current NPA classification: the Chinese Owl shows a higher genomic proximity to the Owls & Frills group than to the Structure, while the Egyptian Swift to the Syrian group than to the Form. As for those breeds that are not currently recognized by the NPA, our results demonstrate that the Birmingham Roller and the Backa Tumbler share the highest genomic similarity with the TRHF group, the California Color Pigeon and the Saxon Fairy Swallow with the Color, and the Mindian Fantail with the Structure.

**Abnormal Samples**

According to the phylogeny, the sample Indian-Fantail_03 does not share proportions of individual ancestries with nor is located near the remaining group of Indian Fantails on the MDS. Instead, this sample shares considerable genetic similarity with the Cumulet group, corroborating what was previously seen on the phylogeny and violin plot (Fig. 1 and 2). In the same way, the two Iranian Tumbler samples do not show similar proportions of individual ancestries and are also located on different regions of MDS. Unfortunately, no photo voucher exists for the IranianTumbler_02 sample, thus we cannot fully exclude the possibility of mislabelling issues. However, we highlight that this sample was collected at a breeder’s house, therefore it is plausible that it was an experimental, hence out-crossed bird.
Supplementary Figures

Supplementary Fig. 1 | Schematic example of a GBS chimeric read.

<table>
<thead>
<tr>
<th>Read section mapped to scaffold ScoHet5_2944_97651</th>
<th>Cut-site</th>
<th>Read section mapped to scaffold ScoHet5_2296_836492</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGCATTAAGCACCTCGTGCTAGACTTTGGATTTTTGACATGCATTGTCCCATAAAGAAAACAAATTGCTGCTCTGAGTACCCTTAAACA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GBS Chimeric Read
Supplementary Fig. 2 | Heatmap based on the presence/absence matrix. Columns represent individual samples, while rows represent clusters of loci. Those samples that failed to produce sufficient GBS reads and blanks samples are marked in red.
Supplementary Fig. 3 | Global Depth (GD) density plots. (a) for all the 50 WGS samples that passed our initial filter. (b) for all samples that passed our initial filter (in red). A subset of this data only containing the possible paralog loci is also plotted (in light blue).
Supplementary Fig. 4 | Examples of phenotypic diversity among pigeon breeds. (A) Stray pigeon presenting the blue-bar ancestor morph. (B) Hamburg Sticken pigeon presenting the HeadCrest, ChestFrill and ReducedBeak traits. (C) West of England Tumbler pigeon presenting the FootFeathering trait. (D) Pomeranian Pouter pigeon presenting the InflatedCrop trait. (E) Scandaroon pigeon presenting the EnlargedBeak trait. (F) English Carrier pigeon presenting the ProminentWattles trait. (G) Old Dutch Capuchine pigeon presenting the HeadCrest trait (Hood). (H) Barb pigeon presenting the ProminentWattles trait. (I) African Owl pigeon presenting the ReducedBeak trait. (J) Figurita pigeon presenting the ReducedBeak and the ChestFrill traits. (K) Fantail pigeon presenting the ExtraTailFeathers trait. (L) Laugher pigeon representing a breed that presents the SpecialVoice trait. The photos presented in A, E, F, G, H, K, L were taken by Hein Van Grouw, while the ones presented in B, C, D, I and J were taken by Michael D. Shapiro.
**Supplementary Fig. 5 | Violin plot of genetic distances.** Coloured dots represent pairwise comparisons between IndianFantail_03 and all other Indian Fantails (red dots; Intra-breeds), IndianFantail_03 and all Cumulets (red dots; Inter-breeds), Laugher_01 and Shakhsharli_01 (blue dot), Fantail_02 and Fantail_09 (orange dot), Scandaroon_01 and Scandaroon_02 (brown dot), all Jacobins and all Old Dutch Capuchines (purple dots), EnglishLongFacedTumbler_01 and all Parlor Rollers (light green dots) and Laugher_01 and MarcheneroPouter_01 (dark green dot). The width of each plot is proportional to the number of samples at a given genetic distance.
Supplementary Fig. 6 | Proportion of observed heterozygous sites. (A) plotted per NPA group. (B) plotted per breed. All absolute values can be found in the Supplementary Spreadsheet.
Supplementary Fig. 7 | Sites and SNPs information. (A) Regression analysis plot showing the correlation between the scaffold sizes and number of sites found in each scaffold. (B) Density plot comparing the distances among cut-sites (in red) and among SNPs (in light blue).
Supplementary Fig. 8 | Multidimensional Scaling Analysis (MDS). Dimensions 2 and 3 of are plotted and each point on the plot represent a single individual.
A)
### Drift Parameter

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td></td>
</tr>
</tbody>
</table>

### Supplementary Information

Pacheco et al., 2019

Migration

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

### Natural History Museum of Demark

Article Under Revision

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Authors

PhD Thesis | George Pacheco
**Supplementary Fig. 9 | TreeMix phylogenies for pigeon breeds and model residuals.** Model scenarios with (a) no migration edge, (b) one migration edge, (c) two migration edges, (d) three migration edges, (e) four migration edges, and (f) residuals of the five-migration model presented in Fig. 5. The out-group is marked in red in all phylogenetic plots.
Supplementary Fig. 10 | Linkage-disequilibrium (LD) across pigeon breeds. Decay of LD, as measured by \( r^2 \), across physical distance. Pairwise LD was binned into 200-bp bins, and the 0.95 quantile used to infer the best-fitted curve (solid-line).
Supplementary Fig. 11 | Manhattan plot for the GWAS analysis for the HeadCrest trait. The green line indicates the permutation derived significance threshold for association (P < 1.91 X 10^-8). The genomic coordinates of SNPs with significant associations to the phenotype are specified. The Y-axis shows the –log10-transformed two-tailed P-value of each SNP from the GWAS meta-analysis (of linear and logistic regression statistics), while the X-axis shows base-pair positions along the scaffolds.
Supplementary Spreadsheet

It can be readily consulted online on Google Docs or downloaded from GitHub.

Bioinformatic Documentation

It can be readily downloaded from GitHub.
Chapter II

I shall say that I found the project presented in this chapter the most challenging of all. It is true that part of its dataset was generated during my MSc, however it was only during my PhD that I truly commenced acknowledging its full extension. This project led me to the realization that leading a scientific study with more than 30 co-authors based in 20 different countries or so is no trivial task. Neither is the endeavour to amass biological samples from across the globe even if the relevant species is a not threatened animal such as the ordinary pigeon. We apply here the same bioinformatic approach described in Chapter I aiming to not only better understand the phylogenetic relationships amongst the current pigeon feral populations, but also to attempt to elucidate how these populations of urban birds first came into existence. Importantly, our conclusion that some localities can still harbour undisturbed rock dove colonies despite the long history of feral expansion has crucial relevance for the conservation of the will C. livia entity.
On the Origin and Spread of Feral Pigeons

George Pacheco1, Filipe Garrett Vieira1, Michael D. Martin2, Morten Tange Olsen1, Michael D. Shapiro3, Pavel Hulva3, Tânia de Freitas Raso, Peter Njoroge, Concepción Salaberría, Isabel López-Rull, Carles Lalueza-Fox, Oscar Ramírez, María C. Ávila-Arcos, Patricia Rosas Escobar, Rui Faria, Miguel Carneiro, Graciela Sotelo, Jóhannis Danielsen, Nizar Haddad, Fares Khoury, Roi Dor, Ali Halajian, María Belén Arias, Oliver Krone, Susanne Auls, Sampath S. Seneviratne, Kajanka Mathiaparanam, Michael Bunce, Megan L. Coghlan, Jon Fjeldså & M. Thomas P. Gilbert1,3

1Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark.
2NTNU University Museum, Norwegian University of Science and Technology (NTNU), NO-7491, Trondheim, Norway.
3School of Biological Sciences, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112, USA.

Department of Biology and Ecology, University of Ostrava, Ostrava 71000, Czech Republic.

Faculdade de Medicina Veterinária e Zootecnia, Universidade de São Paulo, Departamento de Patologia, São Paulo, Brazil.

Ornithology Section, Department of Zoology, National Museums of Kenya, Museum Hill Road, P.O. Box 40658, P.C. 00100, Nairobi, Kenya.

Centro de Investigación en Ecosistemas, Universidad Nacional Autonoma de Mexico, Michoacan, Mexico.

Departamento de Ecologia Evolutiva, Museo Nacional de Ciencias Naturales, Madrid, Spain.

Avian Evolution Node, Department of Zoology and Environment Sciences, University of Colombo, P.O. Box 1490, Colombo 03, Sri Lanka.

Institute of Evolutionary Biology, Universitat Pompeu Fabra – Consejo Superior de Investigaciones Científicas, Doctor Aiguader 88, 08003 Barcelona, Spain.

Department of Animal and Plant Sciences, University of Sheffield, Alfred Denny Building, Western Bank, S102TN Sheffield, UK.

CIBIO/InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, Universidade do Porto, 4485-661, Vairão, Portugal.

IBE, Institute of Evolutionary Biology (CSIC-UPF), Department of Experimental and Health Sciences, Pompeu Fabra University, Doctor Aiguader 88, 08003 Barcelona, Spain.

Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n. 4169-007 Porto, Portugal.

University of the Faroe Islands, Faroe Islands.

National Center for Agricultural Research and Extension, 19381 Baqa’, Jordan.

Department of Biology and Biotechnology, American University of Madaba, Madaba, Jordan.

Department of Zoology, Tel Aviv University, 6997801 Tel Aviv, Israel.

Natural History Museum, Imperial College of London, Cromwell Road, Kensington, SW7 5BD London, United Kingdom.

Department of Biodiversity, University of Limpopo, Turfloop Campus, Private Bag X1106, 0727 Polokwane, South Africa.

Department of Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Strasse 17, 10315 Berlin, Germany.

Vetgenomics SL, Edifici Eureka, Campus UAB, 08193 Bellaterra, Barcelona, Spain.

Trace and Environmental DNA (TrEnD) Laboratory, Department of Environment and Agriculture, Curtin University, Perth, WA 6102, Australia.
Abstract

The rock pigeon (*Columba livia* Gmelin, 1789) is a species presumably native to Mediterranean, Saharo-Arabian and Eastern Oriental regions. During the Neolithic, its domestication took place somewhere in the Middle East probably via the commensal pathway. Initially, the domestic pigeon served an array of purposes, which ranged from practical ones that included a source of food and fertiliser to spiritual ones. As time progressed its range of uses diversified, including use in religious rituals, and carriers of messages, and within the past few centuries it has also received considerable attention from those interested in generating fancy breeds. At some point during this process the first feral pigeons arose, whose populations subsequently undertook a remarkable expansion, that has resulted in them being found today across almost the entire global urban landscapes. Indeed the spread of these feral birds has been so prolific, that it raises questions about whether any true wild rock pigeon colonies still exist, or whether they have been admixed with, or even fully replaced, by feral birds. While several studies have investigated the complex evolutionary history of pigeon breeds, none have yet addressed the question of pigeon feralisation, and how this evolutionary process might be jeopardizing the species’ status as a wild entity. In this study, we generated and analysed a genomic dataset produced using the Genotyping-by-Sequencing (GBS) method of 450 feral pigeons sampled across 41 worldwide localities. Our analyses show that the worldwide pigeon populations analysed may be divided into four major groups, each with different levels of genetic diversity and contamination with domesticated genotypes. We also find signs of strong population structure, including very divergent clades of what seems to be relatively wild populations. Lastly, we find evidences of human-mediated dispersal through past colonial links.

Keywords: Feral Pigeon, Genotyping-by-Sequencing, Population Genomics
Introduction

Archaeological evidence suggests that the rock pigeon (Columba livia Gmelin, 1789), and in particular the C. l. livia subspecies, was first domesticated during the Neolithic period in the Middle East, probably via commensal pathway\(^1\), being initially used as a source of both food and fertiliser, as well as in religious rituals. Later on, the extent of its service to humankind spun a wider variety of roles, including a tool for communication, source of medicine, and even as a navigation aid\(^2\). Furthermore, in addition to its practical functional roles, and in parallel with many other domestic animals such as dogs, chickens and cats, the eighteenth century witnessed an explosion of interest in the development of so-called fancy breeds. Such interest led to the establishment of numerous pigeon breeds, of which over 230 are currently recognised by the American National Pigeon Association (NPA; www.npausa.com). This collection of breeds encompasses a truly fabulous amount of phenotypic diversity, which has long attracted the attention of scholars, and even formed a cornerstone of Darwin’s nascent thoughts on his famous theory concerning the evolutionary processes\(^3\).

The history of the pigeon’s domestication has also been tightly coupled with a correlative evolutionary process—feralisation. Domesticated pigeons are ideally suited to occupying human-made environments, given the natural behaviour of their wild relatives for inhabiting cliff ledges. Thus the process of feralisation was almost certainly initially driven through domestic pigeons escaping from captive stocks (kept within Europe, North Africa and Western Asia) who then successfully populated the immediate urban areas. As with other feralised domesticates, the species experienced an extreme ecological range expansion much later in time, when during the colonial period they were transmitted to, and subsequent release across, almost all continents of the world. Therefore, at present, the feral pigeon is ubiquitous across the world’s urban landscapes, where it is often considered a pest species requiring active management. The history of rock pigeons however, is not all about success. Given that both domestic and feral pigeons can almost certainly still interbreed with their wild ancestor, it has been proposed that in regions of co-occurrence, the wild rock pigeon gene pools (and thus their integrity as a natural species) might be to some extent contaminated with domestic genotypes as has been demonstrated for other domestic species\(^4,5\).

Although studies aiming to shed light on the genomic relationships among pigeon breeds have been conducted\(^6,7\) (Pacheco et al., 2019), fundamental questions concerning the current populations of feral pigeons have yet to be addressed in depth. These include i) which were the breeds that contributed most to the formation of the worldwide extant feral pigeon populations, and ii) which are the genomic relationships among these populations. In addition, in light of the accelerated global expansion of domestic stocks and feral populations throughout the most recent centuries, a third question of utmost relevance is iii) whether the wild rock pigeon has undergone a process of genomic extinction as seen with the wild ancestors of other domesticates\(^5,8\).

In this study, we employed the Genotyping-by-Sequencing (GBS) method to generate a genomic dataset for 450 free-living pigeons from 41 localities covering a worldwide distribution. We use this dataset to reconstruct the phylogenetic relationships among these populations, as well as to investigate their patterns of both genetic structure and diversity revealed by analyses of MDS, Admixture and population genetics statistics. To the best of
our knowledge, this is the most extensive study on a feral animal hitherto performed with regards to sampling effort.
Results

Sampling Effort

In order to investigate the genomic patterns of current pigeon populations of different evolutionary histories, we intentionally divided our sampling effort into four categories: i) remote localities within the assumed natural range: Crete (Greece), Eja, Kunoy, Ljos Air, Noldsoy, Sumba, Torshavn (Faroe Islands), Hadi Hidan (Jordan), Sardinia (Italy), Vernelle (France); ii) urban localities within the assumed natural range: Abadeh, Isfahan, Lahijan, Nowshahr, Tehran (Iran), Barcelona (Spain), Berlin (Germany), Cambridge, London (England), Copenhagen (Denmark), Guimarães, Lisbon (Portugal) Jihlava, Prague (Czech Republic) and Tel Aviv (Israel); iii) urban localities outside the natural range: Denver, Salt Lake City (USA), Johannesburg (South Africa), Mexico City, Monterrey, San Cristobal de las Casas, Tlaxcala de Xicohtencatl (Mexico), Nairobi (Kenya), Perth (Australia), Salvador, Tatuí (Brazil), Santiago (Chile); iv) captive population claimed to be derived from wild pigeons: Tel Aviv Colony (Israel).

Furthermore, to help root the evolutionary relationships between these groups, we also included a small number of individuals representing the *Columba livia intermedia* (Strickland, 1844) subspecies. Specifically in this regard we sampled five populations from Sri Lanka, where two of them were from urban localities (Colombo and Trincomalee), one was from a Conservation National Park (Pigeon Island) and two others were captive populations maintained by local breeders (Wattala and Wellawatte) (Supplementary Fig. 1). Moreover, in order to check for data reproducibility, we sequenced two of the samples twice (Tehran_16-GBS and Perth_02-GBS) to serve as replicates. Finally, to serve as external outgroups, we also collected five samples of *Columba palumbus* (Linnaeus, 1758) in Copenhagen (Denmark), one captive sample of *Streptopelia risoria* (Linnaeus, 1758), and the same *Columba rupestris* sample that was previously fully sequenced$^6$ (we also included the WGS library to serve as another replicate) (Supplementary Spreadsheet).

We hypothesize that populations inhabiting remote localities within the believed natural range will show the lowest levels of contamination with domestic genotypes, while those populations inhabiting urban localities within the believed natural range will have moderate signals of admixture with domesticates. In addition, those populations outside the natural range will show the highest levels of influence by domestic genotypes, and, given that these populations were formed through human-mediated dispersals, we also expect that there will be a signal linking former colonies-colonizers relationships (e.g. London and Johannesburg).

Sequencing Data and Filtering

We generated 1,062,089,639 reads using the (GBS) method$^4$ from all 488 samples. The percentage of GBS chimeric reads removed ranged from 0.034% to 4.65% (mean of 0.69%) per sample. Based on the presence/absence heatap, we excluded 10 samples alongside the 6 blanks from further analyses (Supplementary Fig. 2). The sequencing depths across the regions of interest ranged from 0.069X to 19.97X (mean of 2.32X) (Supplementary Spreadsheet). This analysis shows that there were large discrepancies in the levels of sequencing effectiveness across all predicted cut-sites, with 68.86% of the loci having no data at all for any sample. This pattern has already been described elsewhere (Pacheco et al., 2018), and we believe that the explanation presented by this previous study would also apply for our current study given the similarity of the datasets of both...
studies. Moreover, based on the Global Depth (GD) results (Supplementary Fig. 3), we run the downstream analyses only considering sites having a mean GD of 150X per sample.

Population Genetics Statistics

Hybridisation between domestic lineages and their respective wild ancestors is an issue of paramount conservation importance as the genomic integrity of wild lineages can be disturbed by contamination with domestic genotypes, either through direct admixture with purebred lineages or through indirect contact with feral stocks\textsuperscript{6,10}. To shed light into the levels of potential domestic contamination, as well as isolation among the sampled populations, we calculate several population genetics estimates for the different pigeon populations as well as F\textsubscript{st} levels between them.

First, we calculated the observed levels of heterozygosity (H\textsubscript{o}) for each sample present in Dataset 1 (all samples). Among the free-living pigeon individuals, the H\textsubscript{o} ranged from 0.1157% to 0.2747% (average of 0.2231%) (Supplementary Spreadsheet), with considerable variation within each population and the occurrence of several outliers (Supplementary Fig. 5). On the other hand, the H\textsubscript{o} ranged from 0.1818% to 0.2389% (average of 0.2176%) among the five C. palumbus samples.

Moreover, we calculated the nucleotide diversity (\(\pi\)), Watterson’s \(\theta\) (\(\theta\textsubscript{w}\)), and Tajima’s D across the pigeon genome for the 35 populations that had five or more individuals (the Crete population was also included due to its relevance), as these genetic estimates only apply for population data (for the 2 replicate samples only the ones with higher coverage were used) (Supplementary Fig. 6). \(\pi\) ranged from 0.001764 to 0.003080 (average: 0.002584; C. palumbus: 0.002427); \(\theta\textsubscript{w}\) ranged from 0.001507 to 0.003168 (average: 0.002483; C. palumbus: 0.002644) and the Tajima’s D ranged from -0.643739 to 0.854540 (average: 0.248505; C. palumbus: -0.413195). Interestingly, the values calculated for the five C. palumbus individuals in our study were lower than expected for a wild species, but we believe that the fact we analysed synanthropic individuals might have prevented us from accessing the true values for this species (Supplementary Spreadsheet).

We next elected to calculate the pairwise F\textsubscript{st} between populations to provide preliminary insights into the relationships between the different populations. A heatmap plotted using this data, after excluding the populations of Tel Aviv Colony and C. palumbus that showed extremely high levels of divergence (Supplementary Fig. 7), clearly shows that three clusters are recovered. The first is contains populations that are very divergent from the rest (Pigeon Island, Trincomalee, Torshavn, Crete, Sardinia, Vernelle, Wadi Hidan, Abadeh and Tehran). The second contains populations from mostly Europe and former colonies (Tel Aviv, Barcelona, Guimaraes, Lisbon, Berlin, Prague, Denver, London, Perth, Salvador, Copenhagen, Johannesburg and Salt Lake City). The third contains populations from Mexico (Santiago, San Cristobal de las Casas, Monterey, Tlaxcala de Xicohtencatl and Mexico City).

Overall, we see a wide range of genetic diversity levels among the sampled populations, similar to both domesticated (e.g. duck breeds: 0.0020 to 0.0028) and wild species (e.g. mallard: 0.0040\textsuperscript{11}). Tajima’s D show a general pattern of positive values, indicating population contractions, except for the populations from Tehran, Abadeh, Wadi Hidan, and Sardinia. Taken together, these results seem to group the current pigeon diversity into
four groups, depending on the levels of genetic diversity (high, intermediate, or low), while the group with high levels of genetic diversity can be further subdivided depending on the Tajima’s D values (positive or negative). These grouping are further supported by the $F_{st}$ distances, that clearly forms groups around the populations with both intermediate and high genetic diversity.

**Phylogenetic Relationships Among Pigeon Breeds**

Even though the levels of genetic diversity can give initial valuable insights regarding population’s past history, it is also important to infer the evolutionary relationships among the current pigeon populations. For that, we conducted the first phylogenetic reconstruction for non-domestic pigeon populations. Specifically, we first performed a Neighbour-joining (NJ) phylogenetic analyses based Dataset 1 in order to confirm the appropriateness of the outgroups available (Supplementary Fig. 8). Having confirmed that *Columba rupestris* is the most appropriate outgroup, we proceeded with a Maximum-likelihood (ML) phylogenetic analysis based on Dataset 2 (excluding *S. risoria* and *C. palumbus*). Overall, the ML phylogeny reconstructed reveals strong phylogenetic structure across the different pigeon populations studied (Fig. 1). Despite the low bootstrap values (BS) recovered for the internal nodes (probably because the evolution of pigeons is not a truly bifurcating process), we delineated six main clades on our phylogeny based on geographical distribution and biological assumptions; Clade A: The Pigeon Island and Trincomalee *C. l. intermedia* populations; Clade B: Abadeh, Tehran, Crete, Sardinia, Vernelle and Faroe Islands; Clade C: Tel Aviv, Tel Aviv Colony and Wadi Hidan populations; Clade D: Nairobi, Colombo, Lahijan, Nowshahr, Wellawatte and Isfahan populations; Clade E: Guimaraes, Barcelona, Lisbon, Salvador, Tatui, Denver, Santiago, Tlaxcala de Xicohtencatl, Mexico City, Monterrey and San Cristobal de Las Casas populations; Clade F: Jihlava, Prague, Berlin, Salt Lake City, Johannesburg, London, Cambridge, Perth and Copenhagen. Interestingly, the four individuals belonging to the Wattala population do not form a monophyletic clade and are found spread across the phylogeny.

It is noteworthy that Clade A, as well as both the Faroe Islands and the Tel Aviv Colony populations, have 100% of BS, with the former also having a very elongated branch length. Moreover, all the three replicates (namely, *Crupesrstr_01*, *Tehran_16* and *Perth_02*) behave as expected, all clustering with 100% of BS. Furthermore, although both the eight Wattala and Wellawatte captive populations were located within the Colombo region, these populations do not form a distinct clade together with the Colombo population with some Wattala individuals seating on a rather different region of the phylogeny. As for the three captive populations included in our study, all the 6 individuals belonging to the Tel Aviv Colony population form a clade with 100% of BS and phylogenetic affinity with the Tel Aviv and the Hadi Hidan populations. The branch of this cluster is the longest across the entire phylogeny, in agreement with the fact that this colony has been maintained inbred for several generations.

**Population Structure Among Pigeon Populations**

Although the phylogeny gives some interesting preliminary insights, admixture probably played a critical role in the formation of current pigeon populations. In order to shed light on their genetic composition and
distribution, we performed several analyses (Admixture, MDS, and TESS3R) more suitable in situations where extensive gene flow is expected. With the aim to access the patterns of population structure among the pigeon populations present in our study, we subsequently conducted both Admixture and Multidimensional Scaling (MDS) analyses based on Dataset 3 (excluding S. risoria, C. palumbus, C. rupestris, Replicates, Wattala and Wellawatte) (Fig. 2-3). Both show an overall population structure across the studied pigeon populations, and we can highlight three main clusters: i) one very divergent from the rest of populations, which separates along MDS dimension 1 and splits at K=2 (Pigeon Island and Trincomalee); ii) a second including geographically distant samples, which separates along MDS dimension 2 and splits at K=3 (Faroe Islands, Crete, Sardinia, Vernelle, Wadi Hidan, Abadeh and Tehran); ii) and a third cluster that splits at K=4 (Colombo, Lahijan, Nairobi, Nowshahr and Isfahan). The remaining populations seem to be more closely related and form clusters at much higher Ks. For instances, i) all four Mexican populations plus the Santiago population form a unique cluster at K=18; ii) Berlin and Prague at K=14; iii) Salt Lake City, Virginia, Copenhagen (even though it shares some ancestry with the previous cluster), all the English individuals, the Johannesburg and the Perth populations still form a relatively well preserved cluster at K=19. Some populations seem to be quite admixed (Denver, Tel Aviv and Guimaraes) while, on the other hand, other populations are assigned to their own clusters.

In order to better visualise our Admixture results, we used tess3 to plot the results for K=6 on a map of Europe and Middle East (Fig. 4). This plot shows that there seems to be two main components in Europe: one including the pigeons from urban regions (Lisbon, Guimaraes, Barcelona, Cambridge, London, Copenhagen, Berlin, Prague and Jihlava), while another mainly encompasses individuals from isolated localities, such as the Faroe Islands, Sardinia, Vernelle and Crete. In addition, a third cluster is found in the Middle East encompassing both the populations from Tel Aviv and Wadi Hidan, while a fourth is found grouping all three Iranian populations (except for Tehran and Abadeh that group with the isolated European populations).

**Contribution of Pigeon Breeds to Current Non-domesticated Populations**

Three competing hypotheses regarding the origin of the current urban pigeon populations within the rock pigeon native range can be considered. One would be that these populations are the direct descendants of the wild populations that, with the extension of human settlements, have become synanthropic. The second, that current populations are formed by long-standing events of introgression between domestic (escaped individuals from captivity) and wild stocks; and a third where these escaped individuals have completely displaced the original wild pigeon colonies and are thus direct descendants of purebred individuals. In this way, we hypothesize that populations located in more urbanized regions would have higher levels of domestic backgrounds, while those populations located in more remote regions would have lower levels of domestic genomic contribution and hence their genomes would virtually resemble those of wild pigeons. Moreover, previous studies have suggested that the Racing Homer breed would be the breed to most contribute to the formation of feral populations given the high numbers in which it is raised.

Thus, in order to shed light on these hypotheses, we performed a supervised admixture analysis on Dataset 3, with the addition of 11 Racing Homer individuals$^{6,12}$ (Pacheco et al., 2019) under the assumption that
these are composed of a single ancestral component (Supplementary Fig. 9). Our results show that pigeon populations on the rock pigeon wild range can be grouped into three groups depending on the different levels of contamination with domestic genotypes: high (Copenhagen, London, Cambridge, Berlin, Prague, Jihlava), low (Lisbon, Guimaraes, Barcelona, Tel Aviv), and almost absent (Faroe Islands, Sardinia, Crete, Vernelle, Tehran, Nowshahr, Lahijan, Isfahan, Abadeh and Wadi Hidan).

**Gene Flow between Pigeon Populations**

Convoluted evolutionary histories that are rife with inter-lineages crosses, such as that which gave rise to the current pigeon populations, inevitably give non-bifurcating phylogenies, which cannot be suitably reconstructed through traditional phylogenetic methods\textsuperscript{13,14}. Thus, based on Dataset 4, we used the software TreeMix to reconstruct the phylogenetic affinities among pigeon breeds while taking into account events of gene-flow among these populations.

The inferred migration events (m=1 to 5) do not seem to be consistent across the analysis (Fig. 5). The first migration event (m=1) occurs between the branch shared by Nowshahr and Isfahan and the node shared between Lahijan and Colombo populations. When allowing for 2 migrations (m=2), the previous migration disappears, and one appears from the branch of Isfahan to the Tehran population, and another from the basal branch of all samples (except Pigeon Island and Trincomalee) to the node shared between Lahijan and Colombo populations. When setting m=3, one migration occurs between the branch shared by Nowshahr and Isfahan to the Tehran population, a second from the basal branch of all samples (except Pigeon Island and Trincomalee) to the node shared between Lahijan and Colombo populations. When m=4, one migration occurs from the branch of the Wadi Hidan population to the Tel Aviv population, a second from the branch of the Isfahan population to the Tehran population, a third between the basal branch of all samples (except Pigeon Island and Trincomalee) to the node shared between Lahijan and Colombo populations and a fourth from the basal branch of all samples to the Nairobi population. Finally, when allowing for 5 migrations (m=5), one occurs between the Wadi Hidan and the Tel Aviv populations, a second between the branch of the Isfahan population to the Tehran population, a third between the basal branch of all populations (except Pigeon Island and Trincomalee) to the node shared between Lahijan and Colombo populations, a fourth between the basal branch of all populations (except Pigeon Island and Trincomalee) to the Nairobi population and a fifth between the branch of the Trincomalee population to the Colombo population (Supplementary Fig. 10).

**Effective Migration Surface**

In order to better visualise the complex relationships between the pigeon populations’ genetic and geographic patterns, we estimated the effective migration surface on the Europe and Middle East region, using the EEMS software based on Dataset 3. The EEMS results (Fig. 6) show that feral pigeon populations from Northern Europe (London and Copenhagen) and Iran (except Tehran and Abadeh) tend to have lower than average levels of $\pi$ and, at the same time, that all populations seem to be relatively isolated, with little signs of
gene flow between them. Nonetheless, these results should be taken with caution since the low density of sampling points might not provide enough resolution to correctly infer levels \( \pi \) and rates of migration.

**Genome-Wide Association Study**

It is noteworthy to notice that although introgressions between wild and domestic lineages are generally considered to be maladaptive, it is also known to favour new adaptations\(^1\). Furthermore, it has been argued that urban feral population are under specific selective pressures that can lead to even speciation, which would give us an unique opportunity to study organic evolution in action\(^16,17\). Moreover, it has been shown that it is possible to perform Genome-Wide Association Study (GWAS) based on GBS data despite some limitations imposed by the characteristic of this sort of data (Pacheco et al., 2019).

Thus, in an attempt to identify potential genomic regions that could be linked with feralisation, we performed a GWAS on Dataset 4, scoring each population as feral or non-feral (Supplementary Spreadsheet). Unfortunately, the GWAS performed did not identify any genomic regions as being significantly associated with the feral phenotype (Supplementary Fig. 11).
Discussion

Prior to its domestication during the Neolithic period, somewhere in the Mediterranean region\textsuperscript{18}, the rock pigeon (Columba livia Gmelin, 1789) is believed to have inhabited a vast natural range covering geographical locations as distant as the cliffs of southern Norway, the shores of the Mediterranean and the coastline of the Indian Ocean \textsuperscript{18,20}. Furthermore, nine morphologically different subspecies are currently recognised for the species. Nonetheless, given the rapid and constant spread of feral populations worldwide, it is reasonable to hypothesize that wild colonies have been threatened by admixture with sympatric feral populations, which could jeopardize the rock pigeon’s conservation as a wild species. Therefore, even though the conservation status of this species is frequently considered of least concern, there is an ongoing debate as to the importance of measuring the degrees of admixture with domestic and/or feral genotypes, which may guide future conservation efforts.

When the results of all our analyses are taken together, our results indicate that all the investigated populations can be grouped into four major groups. The first one, being closest to the outgroup, is formed by individuals from both Pigeon Islands and Trincomalee populations (Clade A). The populations in this group have high values of $H_o$ and $\pi$ (similar to wild mallard populations\textsuperscript{11}) as well as positive values of Tajima’s D (hinting at population contraction). In addition, these populations are the most divergent on MDS dimension 1 and are assigned to a unique Admixture cluster at K=2. This, together with its geographical location (South Asia) and the fact that they are quite divergent from the closely located Colombo population (see below), is consistent with our original hypothesis that these individuals might actually belong to the subspecies *C. livia intermedia*. However, this hypothesis warrants further studies, since the low density of samples precludes us from delving deeper.

Another group that is relatively close to the outgroup is formed by all populations in phylogenetic Clade B, plus the Wadi Hidan population (Clade C). Most of the populations of this group present high values of $H_o$ and $\pi$ (except for the Faroe Islands), which are also at similar levels to the wild mallard populations\textsuperscript{11}. Tajima’s D show both signs of population contraction (Vernelle and the Faroe Islands), and population expansion (Tehran, Abadeh, and Sardinia). The populations belonging to this group are split along the MDS dimension 2 and are all assigned to a common Admixture cluster at K=3. Moreover, it is noteworthy to notice that although these populations are genetically similar, they are geographically well apart from each other. This hints that these populations might be remnants of the wild ancestor populations that were displaced from most of Western and Southern Europe by feralised pigeons. Further supporting this idea, are the results from the supervised admixture analysis that show very little levels of domesticated genotypes on these populations. The only exception might be the Faroe Islands population that, with its low values of $H_o$ and $\pi$ (low diversity; reduced effective population and genetic drift), and high Tajima’s D (bottleneck) seems to be the result of an older founder-effect (e.g. Viking colonization).

Another clear group is composed by populations from Europe and their former colonies (Clade E and Clade F). Even though these individuals cluster together on the phylogeny and MDS, the populations are quite different, with the latter showing on average lower levels of $H_o$ and $\pi$, and higher levels of Tajima’s D, probably due to the colonization bottleneck. The European populations also seem to split broadly into two sub-groups:
South West and North Central. The former shows high levels of $H_o$ and $\pi$, and low but positive Tajima D levels, hinting that these populations might be the result of feralised individuals that experienced some introgression from local native wild populations, with whom they overlap in geographical range. The North populations show high Tajima’s D values (similar to the colonies) and intermediate $H_o$ and $\pi$, suggesting these populations are the result of single feralisation events. This interpretation is further supported by the supervised admixture results, where the Northern and Southern populations show high and low levels (respectively) of domesticated genotypes.

Lastly, the group formed by the populations in the phylogenetic Clade D, in addition with the Tel Aviv population (Clade C), is characterized by low $H_o$ and $\pi$ (except for Colombo; probably due to local introgression with C. livia intermedia), as well as positive Tajima’s D values. However, Tajima’s D values vary significantly as some populations present lower values, probably due to past admixture with wild populations (Lahijan and Nowshahr; similar to those of other cities in Southern Europe), while others have quite extreme values (Nairobi, Isfahan; similar to former-colonies), which hints at possible bottleneck or founder effect (e.g. caused by direct influence of the Muslim Empire). Colombo is probably one of the latter, since it does not overlap with native range of C. livia and its lower Tajima’s D levels are probably due to local introgression with C. livia intermedia. Furthermore, all populations in this group cluster together on the MDS (except for Tel Aviv; which is halfway towards the European populations), and are assigned to a common Admixture cluster at K=4 (except again for the Tel Aviv; which higher Ks indicate to be quite admixed).

Despite the fact that we found overall congruence between our ML (Fig. 1) and TreeMix phylogenies when no hybridization events were allowed (Fig. 5), the inferred migrations were not consistent, and led to inconclusive patterns. Nonetheless, we consider noteworthy to notice that all migration events inferred involve the Middle Eastern populations (Tel Aviv, Colombo, Isfahan, Nairobi, Lahijan, and Nowshahr). Moreover, the EEMS results (Fig. 6) are in agreement with this finding as they hint at very little sign of gene-flow between populations. However, this might be a consequence of the complex pigeon evolutionary history in the Europe/Middle-East region, with confounding evolutionary signals from native wild populations, secondary colonizations on both the native range (Isfahan) and new territory (e.g. Copenhagen, Berlin), and feral populations.

Finally, the fact that we could not identify any genomic loci significantly associated with feralisation might be due to some non-exclusive reasons. For instance, the scoring as wild and feral populations might need to be re-evaluated. Also, even though it has been shown that GWAS can be successfully performed based on reduced-representation library sequencing data (Pacheco et al., 2019), this kind of data is not the most ideal for GWAS due to its remarkably low SNP density. If so, the reason it worked on previous studies might be related to the high LD levels of pigeon breeds.
Conclusion

During recent decades, domestic organisms have received great attention from the genomic community, which has facilitated the genome assembly of several domestic species. This increase in genomic data has facilitated numerous studies that aim to correlate specific genomic modifications with a great wealth of phenotypic variations, which are a characteristic outcome of the evolutionary process by means of domestication. On the other hand, however, much less attention has been given to a correlate evolutionary process: feralisation.

With an aim to expand our knowledge of the genomic consequences of feralisation, in this study we analysed 450 free-living pigeons, belonging to 41 populations distributed worldwide. Taking into consideration all our results, we believe that the worldwide populations of free-living pigeons analysed may be divided into four major groups. One formed by the Sri Lankan populations of *C. livia intermedia* (Pigeon Island and Trincomalee), that seem to be virtually free of contact with feral and/or domestic genotypes. The second and third groups include populations from both Europe and the Middle East, but with different degrees of contamination with domestic genotypes. One, composed of populations located in isolated regions (e.g. the Faroe Islands, Crete, Vernelle and Sardinia) that still tend to preserve the genomic background of *C. livia livia*, and another with populations located in urbanised locations (e.g. Tel Aviv, Barcelona, Berlin and Copenhagen) that tend to show fair degrees of contamination. Lastly, a third group would include all populations located outside the rock pigeon’s native range. In general, these populations are genomically marked by a severe founder effect (e.g. Tatui and Nairobi), which is characteristic of human-mediated dispersals (e.g. colonization).

Additionally, even though our results point to the lack of gene flow between nearby populations, we believe that future studies could attempt to measure the levels of migration in a finer scale. For instance, it would be interesting to investigate whether there is any gene flow amongst several adjacent populations on Sri Lanka or amongst several colonies occupying different islands on the Faroe Islands. Furthermore, despite the fact that we could not identify any genomic region correlated with the feral phenotype, we envisage the promising potential use of these feral populations in future studies. For instance, based on a large with full-genome dataset, it should be possible to identify loci not only involved on feral trait, but also those that might have been recently selected due to local adaptations (e.g. adaptation to altitude and extreme climates).
Methods

**Sequencing Data Generation and Processing**

We amassed a worldwide panel of 473 biological samples of *C. livia* through international collaboration. Whenever possible, a voucher photo was taken from both dorsal and ventral orientations in order to record the general plumage pattern of each bird. Both storage condition and method of DNA extraction varied across samples (Supplementary Spreadsheet). We performed DNA quantifications, molecular integrity checks followed by the Genotyping-by-Sequencing (GBS) method⁹ (all samples were sequenced in 6 different lanes). Data demultiplexing, removal of failed samples, removal of possible paralog loci and mapping (including the same regions of interest) was performed as described in (Pacheco et al., 2019).

**Data Analysis**

We produced several datasets using the package ANGSD-v0.921²¹, each including different samples and following different conditions to be used in specific analysis (Supplementary Spreadsheet). Despite the fact that each of these dataset had their own specificities, all of them were generated obeying the same common parameters and conditions previously applied on a similar dataset (Pacheco et al., 2019). Dataset 1 (all samples): 475 samples, 1'225'204 sites, coverages between 10.3X and 566.7X (mean of 68.8X), missing data between 0% to 23.2% (mean of 0.75%); Dataset 2 (excluding *S. risoria* and *C. palumbus*): 469 samples, 1'261'881 sites, coverages between 10.3X and 560.9X (mean of 68X), missing data between 0% and 22.6% (mean of 0.59%); Dataset 3 (excluding *S. risoria*, *C. palumbus*, *C. rupetris*, Replicates, Wattala and Wellawatte): 457 samples, 20'659 SNPs, coverages between 10X and 586.4X (mean of 71.7X), missing data between 0% and 20.2% (mean of 0.54%); Dataset 4 (excluding *S. risoria*, *C. palumbus*, Replicates, Wattala and Wellawatte): 458 samples, 20'705 SNPs, coverages between 10X and 586.5X (mean of 71.6X), missing data between 0% and 20.2% (mean of 0.54%).

To check whether the GBS method was able to produce SNPs equally distributed across the pigeon genome, we performed a regression between the size of each scaffold and the number of sites found in each scaffold. We found a rather strong correlation ($r^2 = 0.99$ and $p$-value $< 2.2e-16$) between scaffold size and number of sites reported (Supplementary Fig. 4), indicating that the GBS protocol successfully yielded sites well-distributed throughout the pigeon genome.

**Population Genetics Statistics**

We followed the instructions provided by ANGSD v0.921²¹ to calculate the unfolded global estimate of the Site Frequency Spectrum (SFS) in order to calculate the observed fraction of heterozygous sites ($H_o$) per sample, the estimates of nucleotide diversity ($\pi$), Watterson’s $\Theta$ ($\Theta_w$) and Tajima’s D (per population), as well as the pairwise $F_{st}$ (per pair of population)²². Specifically, the observed fraction of heterozygous sites ($H_o$; calculated as the ratio between the number of heterozygotes and the total number of sites with information in percentage).
Phylogenetic Reconstruction

For the Maximum-likelihood (ML) phylogenetic reconstruction, we used RAxML-NG v0.5.1b (https://github.com/amkozlov/raxml-ng) to perform two phylogenetic searches using as starting topology either a NJ phylogeny or 20 random topologies. The Neighbour-joining (NJ) phylogenetic reconstruction was based on a pairwise genetic distances matrix calculated directly from the genotype likelihoods outputted by ANGSD using the software ngsDist v1.0.2\textsuperscript{23} with pairwise deletion (--pairwise_del), and inferred using the software FastME v2.1.5\textsuperscript{24} with the SPR tree topology improvement (-s). Both these searches employed the GTR model with discrete GAMMA with 4 categories, mean category rates and ML estimate of alpha (--model GTR+G), as well as used the site repeats optimization option (--site-repeats on). The phylogeny with the highest likelihood was chosen (that incidentally turned out to be the one starting from the NJ phylogeny) and used RAxML-NG to calculate bootstrap values using the bootstrap option based on 100 replicates (--bs-trees 100) and the same setup model used to compute the main phylogeny. We used the online software iTOL v4.0.3\textsuperscript{25} to visualised and plot the final bootstrapped phylogeny.

Inference of Population Structure

We used a pairwise genetic distances matrix to perform a Multidimensional Scaling (MDS) analyses using the R package cmdscale. The software ngsAdmix v32\textsuperscript{26} was used to estimate proportions of individual ancestries for K = 2 up to K = 20 in 100 replicates using default parameters, except for tolerance for convergence (-tol 1e-6), log likelihood difference in 50 iterations (-tolLike 50 1e-3), and maximum number of EM iterations (-maxiter 10000). Lastly, we used TESS3 v1.1.0\textsuperscript{27, 28} software to geographically constrain and plot the Admixture results for K=6 onto the world map.

Contribution of Pigeon Breeds to Current Non-domesticated Populations

We used a variant version of software ngsAdmix v32\textsuperscript{26} (https://github.com/fgvieira/ngsAdmix) to run a supervised analysis, assuming that all Racing Homer individuals are composed of a single ancestral component (in representation of domestic genotypes, since it is the only breed more likely to survive as escaped pigeon). We estimated proportions of individual ancestries from K=2 (one free component) to K=6 (five free components), using the same parameters as above.

Inference of Migration Events

We ran TreeMix v1.13\textsuperscript{14} using default parameters, except for size of block for estimation of covariance matrix (-k 100), sample size correction (-noss), round of global rearrangements after adding all populations (-global), and setting the Crupetsris_01-WGS samples as the out-group (-root Crupetsris_01-WGS). Migration edges were added until residuals did not appreciably decrease (five in our case). The results were plotted using the R function plotting\_funcs provided by TreeMix.
Estimated Effective Migration Surface

To estimate the effective migration surface, we used the software EEMS v20180406\textsuperscript{29}, setting the number of demes to 1000 and defining the outer boundary of the region by the polygon (in latitude–longitude coordinates) \{(138.601, -34.929), (18.4239, -33.9253), (-73.0514, -36.8282), (-122.417, 37.7833), (-93.2667, 44.9833), (-21.8193, 64.1227), (25.4667, 65.0167), (102.6, 17.9667), (138.601, -34.929)\}. We ran the Markov chain Monte Carlo with one million burn-in and two million regular iterations, thinned to one in ten thousand.

Genome-Wide Association Study

From the genotype likelihoods computed by ANGSD, we calculated the expected genotypes and used these as allele dosages. Then, we employed the software \textit{GEMMA}-v0.96\textsuperscript{30} to calculate a centered relatedness matrix (-gk 1) excluding SNPs with minor allele frequency below 0.01 (-maf 0.01). Based on our results, we created a phenotype table scoring the feralisation level across the different populations (Supplementary Spreadsheet). Finally, we used the genetic relatedness matrix to fit a linear mixed model in \textit{GEMMA}, and conducted a likelihood ratio test. Moreover, we performed a permutation test using \textit{MVNpermute}\textsuperscript{31} in order to determine an appropriate significance threshold for each of the association analyses by running 100 permutation replicates and then independently running \textit{GEMMA} on each permuted phenotype matrix. The p-values from all 100 \textit{GEMMA} runs were concatenated, and then the 5th percentile p-value, corrected for the total number of tests, was used to draw the lines of significant of each study. Results were plotted using the \textit{R} package \textit{ggman v0.99.0.beta} (https://github.com/drveera/ggman).
References


Funding

This project was funded by Lundbeck Foundation (award R52-5062) and European Research Council (Consolidator grant 681396) granted to MTPG. G.P. was supported by a Danish Government Scholarship and Tuition Fee Waiver Grant provided by the Danish Ministry of Science, Innovation and Higher Education and subsequently by a Ciência Sem Fronteiras Full PhD Abroad Scholarship (Grant Number: 201761/2014-9) provided by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) in Brazil.

Competing Interests

We have no competing interests.

Acknowledgments

We would like to thank our local lab managers Charlotte Hansen, Pernille V. S. Olsen and Tina B. Brand at the Centre for GeoGenetics for their prompt support during the execution of the project. We are grateful to the Cornell University Biotechnology Resource Center for its genotyping services, especially to Sharon E. Mitchell and all lab technicians that worked in this project. Moreover, we deeply thank Gary Jakeman and Kristian Murphy Gregersen for their fieldwork assistance regarding the sampling in England and Denmark, respectively. We also thank Vladimir Orduña for his willingness to let us sample some of the Mexico City pigeons kept at his lab facilities.

Author Contributions


Data Availability

All demultiplexed GBS sequencing data is publicly available at SRA (Project Number: PRJNA495951), as well as additional data uploaded to the University of Copenhagen’s long term storage (https://sid.erda.dk/wsgi-bin/ls.py?share_id=aKqQoJvH4Y).
Figure Legends

Fig. 1 | Maximum-likelihood phylogeny of pigeons. Phylogeny describing the relationships between over 450 pigeons representing 45 sampling localities. The outgroup (C. rupestris) is depicted in red. Nodes with bootstrap values above 70% are marked with green circles. The inner circular lines represent the inverted scale.

Fig. 2 | Admixture proportions of pigeons. Individual are represented by columns, while rows depict the Admixture proportions based on the assumption of different numbers of ancestral populations (K = 2 - 20). Individuals are sorted by sampling locality (light gray upper labels) and grouped per continent (dark gray upper labels).

Fig. 3 | Multidimensional scaling analysis of pigeons. Dimensions 1 and 2 are plotted and each point on the plot represent a single individual. The dashed ellipses represent the rough distribution of the two most homogeneous groups.

Fig. 4 | Ancestral Map of pigeons. Ancestral map based on K=6 ancestral populations.

Fig. 5 | TreeMix maximum-likelihood phylogeny of pigeons. Five migration events among different pigeon populations are represented by arrows on the phylogenetic graph. The scale bar indicates ten times the average s.e. The out-group (C. rupestris) is marked in red. The model residuals are plotted in Supplementary Fig. 10.

Fig. 6 | Ancestral Map of pigeons. Estimated effective migration rates (a) and diversity rates (b) among pigeon populations in Europe and Middle.
Fig. 1
<table>
<thead>
<tr>
<th>Fig. 2</th>
</tr>
</thead>
</table>

---

Chapter II

23
Fig. 3
Fig. 4
Fig. 5
Fig. 6
On the Origin and Spread of Feral Pigeons

George Pacheco1, Filipe Garrett Vieira1, Michael D. Martin2, Morten Tange Olsen1, Michael D. Shapiro3, Pavel Hulva1, Tânia de Freitas Raso, Peter Njoroge, Concepción Salaberry, Isabel López-Rull, Carles Lalueva-Fox, Oscar Ramírez, Maria C. Ávila-Arcos, Patricia Rosas Escobar, Rui Faria, Miguel Carneiro, Graciela Sotelo, Jóhannis Danielsen, Nizar Haddad, Fares Khoury, Roi Dor, Ali Halajian, Maria Belén Arias, Oliver Krone, Susanne Auls, Sampath S. Seneviratne, Kajanka Mathiapiaranam, Michael Bunce, Megan L. Coghlan, Jon Fjeldså & M. Thomas P. Gilbert1,2

*george.pacheco@snm.ku.dk

1Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark.
2NTNU University Museum, Norwegian University of Science and Technology (NTNU), NO-7491, Trondheim, Norway.
3School of Biological Sciences, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112, USA.

Supplementary Information

Chapter II

Table of Contents

Supplementary Fig. 1 | Map showing sampling effort. ................................................................. 2
Supplementary Fig. 2 | Heatmap based on the presence/absence matrix. .............................................. 3
Supplementary Fig. 3 | Global Depth (GD) density plot. ................................................................. 4
Supplementary Fig. 4 | Sites information. .................................................................................. 5
Supplementary Fig. 5 | Proportion of observed heterozygous sites per sampling locality. ...................... 6
Supplementary Fig. 6 | Population genetics estimates per sampling locality. ....................................... 7
Supplementary Fig. 7 | Pairwise Fst heatmap. ........................................................................... 9
Supplementary Fig. 8 | Neighbor-joining phylogeny of pigeon populations. ........................................... 10
Supplementary Fig. 9 | Supervised admixture proportions of pigeons. ............................................. 11
Supplementary Fig. 10 | TreeMix phylogenies of pigeon populations and model residuals. ..................... 12
Supplementary Fig. 11 | Manhattan plot for the GWAS analysis for the Feral trait. .............................. 18
Supplementary Figures

Supplementary Fig. 1 | Map showing sampling effort. Remote localities within the assumed natural range (green), urban localities within the assumed natural range (brown), urban localities outside the assumed natural range (red), captive population claimed to be derived from wild rock pigeons (blue), potential populations of *C. l. intermedia* (purple).
Supplementary Fig. 2 | Heatmap based on the presence/absence matrix. Columns represent individual samples, while rows represent clusters of loci. Those samples that failed to produce sufficient GBS reads and blanks samples are marked in red.
Supplementary Fig. 3 | Global Depth (GD) density plot.
Supplementary Fig. 4 | Sites information. Regression analysis plot showing the correlation between the scaffold sizes and number of sites found in each scaffold.
Supplementary Fig. 5 | Proportion of observed heterozygous sites per sampling locality. All absolute values can be found in the Supplementary Spreadsheet.
Supplementary Information

A) Nucleotide Diversity

B) Watson's Theta
Supplementary Fig. 6 | Population genetics estimates per sampling locality. (A) Nucleotide Diversity ($\pi$). (B) Watterson’s Theta ($\theta_w$). (C) Tajima’s D. All absolute values can be found in the Supplementary Spreadsheet.
Supplementary Fig. 7 | Pairwise $F_{st}$ heatmap. All absolute values can be found in the Supplementary Spreadsheet.
Supplementary Fig. 8 | Neighbor-joining phylogeny of pigeons. Initial phylogeny describing the relationships amongst over 450 pigeons. The identified outgroup are depicted in different colours: *S. risoria* (red), *C. palumbus* (orange) and *C. rupestris* (dark green). The inner circular lines represent the inverted scale.
### Supplementary Fig. 9 | Supervised Admixture proportions of pigeons

Individuals are represented by columns, while rows depict the Admixture supervised proportions based on the assumption of different numbers of ancestral populations ($K = 2 - 6$). Individuals are sorted by sampling locality (light gray upper labels) and grouped per continent (dark gray upper labels).
Supplementary Fig. 10 | TreeMix phylogenies of pigeon populations and model residuals. Model scenarios with (a) no migration edge, (b) one migration edge, (c) two migration edges, (d) three migration edges, (e) four migration edges, and (f) residuals of the five-migration model presented in Fig. 5. The out-group is marked in red in all phylogenetic plots.
**Supplementary Fig. 11 | Manhattan plot for the GWAS analysis for the Feral trait.** The green line indicates the permutation derived significance threshold for association ($P < 2.59 \times 10^{-8}$). The Y-axis shows the $-\log_{10}$-transformed two-tailed $P$-value of each SNP from the GWAS meta-analysis (of linear and logistic regression statistics), while the X-axis shows base-pair positions along the scaffolds.
Supplementary Spreadsheet

It can be readily consulted online on Google Docs or downloaded from GitHub.

Bioinformatic Documentation

It can be readily downloaded from GitHub.
Postface

The Heavily Bearded Man Was Seldom Wrong

Lab work is an extremely repetitive activity. Thus, once one is fully accustomed to a given lab protocol, all the required actions are mechanically performed. During periods of incubation or centrifuging, one’s mind is then left to wander. Having carried out a great deal of lab work myself during my doctoral studies, I certainly experienced many of these moments. It was during one of these occasions that I unintentionally mentalised a pensive Darwin, who was walking to and fro in front of the pigeon loft built in the backyard of his house in Downe. Behind his mental discomfort was a puzzle to which he could not properly present a solution. Although he could indeed empirically observe that some specimens among his pigeons presented derived and intriguing characteristics, such as extra tail feathers, head crests or foot feathering, he was not entirely certain of the ultimate origin of those deviations. Now released from that state of mind-wandering,—for the procedure I would execute next required my conscious attention—I recall thinking that his lack of understanding regarding the question of variability might have been somewhat irksome to Darwin. Allowing myself to continue in those thoughts, I mentally depicted the Victorian naturalist having to scrutinise all that biological variation present on his premises, without knowing about the idiosyncrasies of genomes, let alone of the biological writing system that is particular to them. I then continued to mentally acknowledge that I would soon finish that DNA extraction, proceeding right after with DNA quantification and inspection of its quality in regard to molecular fragmentation. Some weeks from that extraction day, I would be shipping those DNA extracts overseas so they could be sequenced. Genomic data would follow and analyses would take place. Looking back now, it all seems to have been so straightforward. Naturally, we also had to scrutinise the question pigeon evolution in our own way, but what I am trying to allude to here is the fact that throughout our Pigeon Project we at least possessed a minimum understanding of the theoretical framework behind it. Upon closing the flow hood that day, I could not refrain myself from arriving at the conclusion that it might have taken a great deal of courage from Darwin to so devotedly investigate the pigeon matter despite his absent acquaintance with the whys and wherefores of biological variability.

Perhaps not surprisingly, the very same Charles Darwin that was brave enough to scientifically deal with questions that were beyond the scientific knowledge of his time was also audacious to write the following words:

*It has often and confidently been asserted, that man's origin can never be known: but ignorance more frequently begets confidence than does knowledge: it is those who know little, and not those who know much, who so positively assert that this or that problem will never be solved by science.*

In this context, especially considering all the relevant advances that the field of genomics has experienced during the last years, I believe that I—as a young genomicist—do not have any reasons to not confidently trust that the time when we will have deciphered all the genomic encryptions will not eventually become the very present. Thus, I can indeed envisage a time when a long list containing the required genomic
modifications for the intended phenotypic outcomes will be duly known. Perhaps this futuristic world will become a reality around 632 AF. Maybe even just a few decades before that, nay, maybe one or two millennia afterwards. Despite this lack of chronological assertiveness, I genuinely believe that the age of exhaustive genomic understanding will come true—an age when we will have sequenced and fathomed all Earth’s Genomesphere. Added to which, I cannot refrain myself from conceiving that when this age arrives we will be modifying pre-existing genomes and even writing original ones at our whimsical will in an attempt to add our own titles to the shelves of the always-evolving genomic library.

**And The Pigeons Must Remain Aloft**

As for the time being, I have no doubt that the recognition of the multifaceted *C. livia* as an important model for genomics studies will continuously spread among genomicists. For instance, one question that occupied my mind throughout the course of my doctoral studies concerns the frequency of the so-called complex traits. In other words, which would be the percentage of phenotypes being governed by a complex set of genomic instructions in comparison to those of simple genomic architecture. From my perspective, this descriptive question has important implications for our understanding of organic evolution. I envisage that pigeon breeds have a great potential to assist us in approaching this intriguing question considering the reasons attested in Chapter I. In more pragmatic terms, I would thus be an enthusiastic advocate of the idea to make the domestic pigeon the first domestic animal to have all its officially recognised breeds genomically analysed in an attempt to reveal more and more genotype-phenotype correlations of the myriad of traits encompassed by this breeds.

On a different note, as a young genomicist, I have been also intrigued by the long-standing questions of recent local adaptation and speciation. However, I do not fail to acknowledge that the low rate at which evolution usually occurs often imposes relevant impediments to investigations attempting to tackle these evolutionary questions. Despite these difficulties, considering what was exposed in Chapter II, I am strongly inclined to believe that feral pigeons have much to offer with regards to this sort of investigation, and I think that it is reasonable to deem foreseeable that feral pigeons will become in the near future a model for genomic studies on urban evolution.

Moreover, I would believe that the establishment of feral pigeons as a genomic model will help in broadening our understanding of the feralisation process per se, which is still scarce at present. In fact, we have an ongoing project in collaboration with Hein Van Grouw at the British Natural History Museum that aims to shed further light on the pigeon feralisation process through the analyses of historical samples. Specifically, we prepared 65 museum samples for full-genome sequencing which represent most of the *C. livia* recognised subspecies, with some specimens being collected in remote localities during periods as early as the 17th century. We expect that the upcoming results will provide us with fresh insights into the formation and spread of feral pigeons.

All in all, even though some of its fronts still remain to be completed, I hope that the herein presented PhD thesis did right by Darwin and contributed somehow to further elevate the rock pigeon’s scientific relevance.
Parallel Collaborations

During my doctoral studies I had the pleasure to collaborate with different projects in the vein of evolutionary genomics—from tiny Collembolans to culturally evolving killer whales. I am rather confident that these collaborations not only have contributed the most to broaden my education in this recent field, but also led to the expansion and diversification my scientific network. I strongly think that both these aspects will be of sound relevance in my pursuit to reach a position of research independence during the forthcoming years.
Published & In Press

**Ide Population Genomics Project:** This was a collaboration with Mikkel Skovrind, who was a MSc student at the time advised by Professor Peter Rask Møller. I initially assisted Mikkel in the labs regarding the required standards for the GBS protocol. Later on, I also participated in the discussions concerning his data and results as well as in the writing process of the resulted article.


**Vampire Bat Hologenomics Project:** This was a kind of unique collaboration of mine since I happened to be involved in this project after it had already passed the first round of revision as some more experiments were required to test a hypothesis presented. It was great to be part of this study and experience this kind of co-authorship. I hope I did my best to contribute to this study despite my late involvement with it.


**Orcagenomics Project:** I was responsible for processing a certain number of whale samples for this project; some by myself and some accompanied by Dr. Andrew Foote, who is a Visitor Scholar at the University of Bern in Switzerland and leaded the project. I was indeed very pleased to be involved in a study investigating these fascinating mammals.


**Being Finalised**

**Vampire Bat Genomics Project:** I performed some lab work for this project as well as some bioinformatics analysis. It was definitely enriching to be part of this team.


**Perch Genomics Project:** This will be another article fruit of my collaboration with Mikkel Skovrind mentioned above. This project is also based on GBS data that was generated by Mikkel during his MSc studies.
Collembomics Project: I was involved in this project since its very inception as I helped to decide its experimental design and to secure funding for Dr. Nerivânia Godeiro’s visit to our labs through the submission of a Science without Borders grant application. Furthermore, I was responsible to introduce Dr. Godeiro to the molecular lab techniques and routine as well as to the world of bioinformatics. It was quite interesting to assist in the conversion of a classical taxonomist into a molecular taxonomist. I only hope that I was able to give as much as I gained from this experience. This project led to the two article below.


Ongoing Projects

Avian 10K Genomics Project: I am immensely pleased that I could to be part of the scientific enterprise and I will be forever grateful to Tom for his entrustment to my work and for giving me the opening to appreciate the avian clade. It was delightful to process so many unique samples in the labs and to have a break of it during a fieldwork expedition in the African Savanna. In addition, I truly believe that my participation in the I 10KBirdProject Workshop held in Beijing was of extreme importance for my education as a modern and dynamic scientist. In a nutshell, I have sound interest in this consortium and I am absolutely confident that it will be seen in the near future as a milestone in the field of evolutionary genomics.

Seychelles Magpie-Robin Genomics Project: Tom asked me if I was interested in being part of this project given my experience performing DNA extractions on avian material. I assisted Emily Cavill, who is an MSc Student in Biological Sciences at the University of Bristol in England and leader of the project, with her initial stages at the labs. I do expect this work to generate results that will provide scientific support to conservation actions towards this species.

Moleques do Sul Cavy and Greater Guinea Pig Genomics Project: I was invited by Tom to participate in this projected led by the PhD Student Manuel Escalona enrolled at the Pontifícia Universidade Católica do Rio Grande do Sul in Brazil. As Manuel already had experience in the molecular biology lab as well as in basic bioinformatics, I was merely responsible to introduce him to the library build protocol developed in-house and to the requirements for de novo genome generation using long insert sizes. It was definitely very relevant for me to be involved here for what is to be found, among other aspects, could directly influence in the survival of a species from my homeland.

Amphibian and Reptile Genomics Project: I have been assisting Morten Allentoft, who is an Associate Professor at the Natural History Museum of Denmark, with the initial lab steps required to generate the genomes for some interesting amphibian and reptile animals such as the Black Mamba the Bombina toads.
Appendix I

Genomic population structure of freshwater-resident and anadromous ide (*Leuciscus idus*) in north-western Europe


Section for Evolutionary Genomics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark

**Keywords**

Anadromous, genomic population structure, Genotyping-by-Sequencing, *Leuciscus idus*, salinity, teleost.

**Correspondence**

Mikkel Skovrind, Section for Evolutionary Genomics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark.

Tel: +4525538684;
Fax: +4535321010;
E-mail: mikkelskovrind@gmail.com

**Funding Information**

Aage V. Jensens Nature Foundation, (Grant/Award Number: ‘28272’) DFF-MOBILEX, (Grant/Award Number: ‘DFF - 1325-00136’) Villum Fonden

Received: 23 August 2015; Revised: 12 November 2015; Accepted: 23 November 2015

doi: 10.1002/ece3.1909

**Abstract**

Climate change experts largely agree that future climate change and associated rises in oceanic water levels over the upcoming decades, will affect marine salinity levels. The subsequent effects on fish communities in estuarine ecosystems however, are less clear. One species that is likely to become increasingly affected by changes in salinity is the ide (*Leuciscus idus*). The ide is a stenohaline freshwater fish that primarily inhabits rivers, with frequent anadromous behavior when sea salinity does not exceed 15%. Unlike most other anadromous Baltic Sea fish species, the ide has yet to be subjected to large-scale stocking programs, and thus provides an excellent opportunity for studying the natural population structure across the current salinity gradient in the Danish Belts. To explore this, we used Genotyping-by-Sequencing to determine genomic population structure of both freshwater resident and anadromous ide populations in the western Baltic Sea region, and relate the results to the current salinity gradient and the demographic history of ide in the region. The sample sites separate into four clusters, with all anadromous populations in one cluster and the freshwater resident populations in the remaining three. Results demonstrate high level of differentiation between sites hosting freshwater resident populations, but little differentiation among anadromous populations. Thus ide exhibit the genomic population structure of both a typical freshwater species, and a typical anadromous species. In addition to providing a first insight into the population structure of north-western European ide, our data also (1) provide indications of a single illegal introduction by man; (2) suggest limited genetic effects of heavy pollution in the past; and (3) indicate possible historical anadromous behavior in a now isolated freshwater population.

**Introduction**

It is widely accepted that oceanic water levels will rise and lead to changes in salinity as an implication of climate change (ICCP 2014). In some regions salinity will increase (areas with decreased precipitation), while other areas will see decreases (areas with increased precipitation) (Durack and Wijffels 2010). The Baltic Sea is a mega estuary with salinities ranging from 0 to 20% (Janssen et al. 1999) in which the salinity is predicted to decline throughout the 21st century (Neumann 2010; Meier et al. 2012), leading to changes in species distributions (Vuorinen et al. 2015). However, no published models for salinity in the Baltic Sea incorporate the impact of rising sea levels, making any predictions of future salinities uncertain (Andersson et al. 2015). Nevertheless, any changes will have great impact on anadromous populations of any freshwater fish species that spend part of their lifecycles in the brackish Baltic Sea (e.g., pike *Esox lucius*, perch *Perca fluviatilis*, roach *Rutilus rutilus*, and ide *Leuciscus idus*) (Müller and Berg 1982; Engstedt et al. 2010; Skovrind et al. 2013; Rohtla et al. 2015). Currently, in the western part of the Baltic Sea, species such as these migrate into brackish water close to their maximum salinity tolerance. Such migrations are different to those of species such as salmon, as they do not undergo a process of physiological adaptation to sea water (smoltification), but remain freshwater-adapted throughout their life (Marchall and Grossel...
This limits anadromous behaviour of stenohaline freshwater fish to regions within their salinity tolerance, thus future salinity changes will affect the areas in which this anadromous behavior is possible, unless species can adapt to different salinity tolerances. This in turn will impact the connectivity between populations, and the overall population structure of such species.

The maximum tolerated salinity for ide (Fig. 1) is 15% (Penthon 1985; van Beek 1999). Thus anadromous ide populations that today live in the transition zone between the brackish Baltic Sea and the marine North Sea are believed to be living on the edge of their physiological capacity. Clear evidence in support of this comes from observations that influxes of higher salinity oceanic water from the North Sea often kills thousands of anadromous individuals along the east coast of Zealand, Denmark (Carl 2012a). These observations also suggest that ide in this region may be under selection for adaptation to higher salinity tolerances.

Several previous studies have investigated the genetic structure of populations of other anadromous, stenohaline freshwater fishes, including perch Perca fluviatilis, pike Esox lucius and zander Sander lucioperca. Using D-loop mitochondrial markers and microsatellites, these studies indicated genetic differentiation between coastal, anadromous and freshwater populations (Nesbø et al. 1998; Larsen et al. 2005; Säissä et al. 2010). Similar research on ide has been extremely limited to a small number of studies which used allozymes and focused on river stretches far from the sea (Wolter et al. 2003; Zhigileva et al. 2010, 2013). Furthermore, unlike most other anadromous fish species in the Baltic, the ide has not been subject to large-scale stocking programmes, and it thus provides an excellent opportunity for studying the natural population structure of an anadromous fish species across a salinity gradient ranging from optimal, suboptimal to lethal habitats.

Given this, we applied a population genomic approach to identify genetic inter- and intravariability of ide populations in the Baltic Sea-North Sea transition zone in north-western Europe. In particular our analyses focused on the genomic relationship between freshwater residents and anadromous populations, in order to obtain a glimpse into what the future may hold for anadromous freshwater fishes as climate change transforms the salinities of their habitats.

Methods

Sampling and storage

Fin-clips from 95 ide were collected from nine localities in Denmark, Sweden and the Netherlands by anglers and scientists (Table 1, Fig. 2). Samples were taken from live fish that were released alive immediately after sampling, except for a few voucher specimens now stored in the collection of the Natural History Museum of Denmark (Sample ID: ZMUC P265401-02, P265454, and P265897-5988). All samples were stored in 96% ethanol and kept in minus 20°C freezers. The sample sites fall into three categories: (1) Streams running into brackish water with salinities $<$15‰, (2) Streams running into marine water with salinities $>$15‰, and (3) Streams running into freshwater lake that until the 17th century was a marine fjord (any migration to the sea from this lake is today hindered by a physical barrier). Samples were classified as anadromous when taken from streams in category one. In these streams large shoals of migratory ide are annually recorded by the authors or local anglers in the deltas, and the salinity outside the deltas is normally within the tolerated level. Freshwater resident status was given to samples from category two and three.

DNA extraction, quality-, and quantity control

DNA was extracted using Qiagen’s Blood and Tissue kit (Qiagen Ltd., Crawley, UK) according to the manufacturer’s protocol, although with minor modifications to ensure DNA-yields and concentrations needed for subsequent analyses. Specifically, after adding Buffer AL, samples were incubated at 57°C, for 30 min, with 20 sec vortex every 10 min to ensure complete dissolution of the samples. In the final step only 100 μL of the AE elution buffer was added to increase final concentration. The extraction quality of all samples was verified by the
presence of high molecular weight bands on a 2% agarose gel. The DNA concentration was measured for all extractions with Qubit 2.0 (Life Technologies, Gaithersburg, MD).

### Genotyping

Population genomic data were generated using the Genotyping-by-Sequencing (GBS) approach (Elshire et al.)
2011) – a method that is both economical and provides a relatively high output of single nucleotide polymorphisms (SNPs) distributed across the genome. Extracted DNA was processed by the GBS service provided by Cornell University’s Institute of Biotechnology following their standard pipeline (Elshire et al. 2011). Initial sample optimization indicated the six base cutter restriction enzyme EcoT22I (target site ATGCA/T) exhibited effective genome fragmentation, and this was used for the GBS library generation. All samples were sequenced on Illumina HiSeq 2000 technology, using single read 64 bp chemistry (including library barcodes). Raw data are available from NCBI, accession SRR067014.

Initial data analysis used the zebra fish Danio rerio, Cyprinidae genome (NCBI assembly number GRc10) as reference for the Tassel 4.3 pipeline (Bradbury et al. 2007). However, as only ca. 4% of the reads mapped, this approach was abandoned in favor of the UNEAK3 pipeline (Lu et al. 2013). Tags were defined as groups of more than five identical reads in the UMergeTaxaTagCountPlugin. To ensure a minimal amount of false SNPs to be included in the dataset, we set an Error Tolerance Rate (ETR) of 0.01 on the UTagCountToTagPairPlugin, and a minimum minor allele (MAF) frequency of 0.02 on the UMapInfoToHapMapPlugin. For all downstream analyses the SNPs with more than two alleles were removed. Finally, using Plink 1.9 (Purcell et al. 2007) all SNPs with more than 5% missing data were removed.

Summary statistics and genomic population structure
GenoDive 2.0 (Meirmans and Van Tienderen 2004) was used to calculate summary statistics, including the observed frequency of heterozygotes within sampling sites (H_S), the expected frequency of heterozygotes within sites (H_O), the number of alleles, number of effective alleles, fixation index (statistics included were number of alleles, number of effective alleles) (10,000 permutations), with positive results assessed for Hardy-Weinberg equilibrium described as inbreeding coefficient (G_IS) (10,000 permutations), and negative results as meaning heterozygote deficiency, and negative results (average of 322,466) (Fig. S1B). Sequencing depth (reads/tag) per SNP ranged from 5.49 to 7.10, with an average of 6.50 (Fig. S1D). Mean number of reads, mean number of tags and mean reads/tag in the sample sites did not of tags and mean reads/tag in the sample sites did not of 6.50 (Fig. S1D). Mean number of reads, mean number of tags and mean reads/tag in the sample sites did not of 6.50 (Fig. S1D). Mean number of reads, mean number of tags and mean reads/tag in the sample sites did not of 6.50 (Fig. S1D). Mean number of reads, mean number of tags and mean reads/tag in the sample sites did not of 6.50 (Fig. S1D). Mean number of reads, mean number of tags and mean reads/tag in the sample sites did not of 6.50 (Fig. S1D). Mean number of reads, mean number of tags and mean reads/tag in the sample sites did not.
Summary statistics

The effective number of alleles in the total dataset was 1.319 (Table 1) and the effective numbers of alleles at the sample sites were between 1.309 and 1.398, with VID and TRY having the lowest and highest number respectively. The total expected heterozygosity ($H_T$) level was 0.231 and for the sample sites the expected heterozygosity ($H_S$) (gene diversity) ranged from 0.197 in VID to 0.259 in TRY. The observed heterozygosity ($H_O$) for the sample sites was between 0.197 (VID) and 0.259 (TRY). Inbreeding coefficient ($G_{IS}$) of the total dataset was $-0.002$, but not statistically significant. The inbreeding coefficients for the sample sites were between $-0.042$ and 0.033, but only the positive values were statistically significant, with localities KRO, OND, SUS, KOG, and LOD showing a deficit of heterozygotes. For the individual sample sites the inbreeding coefficients ($G_{IS}$) of the total dataset was $-0.002$, but not statistically significant. The pairwise values being between KRO and TRY sample sites were 0.135 and 0.289 (95% CI: 0.178–0.254), except for the KRO-GUD comparison that had an $F_{ST}$ of 0.070. All pairwise $F_{ST}$ comparisons among sites were significantly different ($P > 0.001$), except KOG and TRY that had a very low $F_{ST}$ value of 0.001 ($P = 0.192$).

The cluster analysis implemented in Admixture showed that all individuals sampled from the same site have similar ancestral fractions (Fig. 3). The plot of two ancestral populations ($K = 2$) shows that there was a basic east-west split between either sides of the marine channel known as ‘The Great Belt’ (Storebælt), that separates the two principal Danish islands of Funen (Fyn) and Zealand (Sjælland) (Fig. 2). When plotting three ancestral populations ($K = 3$), OND is the first site to separate from the other western sample sites. The most likely number of ancestral population was four ($K = 4$) (CV error rate = 0.48697) (Fig. S3) with the sample sites divided into (1) a western cluster consisting of KRO in the Netherlands and the two populations VID and GUD on the Jutland peninsula; (2) a cluster consisting of the OND sample site; (3) a cluster consisting of the POL site; and (4) the four eastern low-salinity sample sites SUS, KOG, TRY, and LOD all clustering together (Fig. 2). In the western cluster, VID separates in the plot of five ancestral populations ($K = 5$) and KRO and GUD stay clustered together. This was consistent throughout plots for 5–8 ancestral populations despite VID being geographically in between KRO and GUD. In the eastern cluster sample sites SUS and LOD separate from the other eastern sample sites (KOG and TRY) in the plots for 6 ($K = 6$) and 7 ($K = 7$) ancestral populations respectively. Sample sites KOG and TRY, which are 3 km apart, cluster together for $K = 2$ through $K = 8$.

The plot of eigenvectors show that the eastern sample sites LOD, SUS, KOG, and TRY cluster on top of each other (Figs. 4, 5). Sample sites GUD and KRO also cluster close to, but distinct from VID. Individuals from OND and POL form their own clusters. Eigenvector 2 shows a very pronounced separation of OND from all other sample sites (Figs. 4, 5) while eigenvector 3 shows an especially large separation of POL (Fig. 5). The eigenvalues show that 58 percent of the variation was found in eigenvectors 1, 2, and 3 combined.

Isolation by distance was statistically insignificant for the correlation of $F_{ST}$ with both waterway distances.
(P = 0.364, R^2 = 0.036) and Euclidean geographic distances (P = 0.376, R^2 = 0.008) (Fig. S2). However, both the correlation of genetic similarity with waterway distances (P = 0.003, R^2 = 0.352) and Euclidean geographic distances (P = 0.006, R^2 = 0.397) were significant when not including the KRO sample site.
Discussion

Population history, salinity, and anadromous behavior

The Isolation by Distance (IBD) results reveal that geographic distance cannot explain all the genetic variation among sample sites. Rather, as suggested in studies of other freshwater species (e.g., perch *Perca fluviatilis*, spined loach *Cobitis taenia*, and bullhead *Cottus*) (Nesbø et al. 1999; Kontula and Viinola 2001; Culling et al. 2006), the large scale patterns of population genomic structuring likely resulted from postglacial population histories and migration routes. Indeed, we observed a split between a western and an eastern clade, which likely reflect such major groups with different evolutionary histories.

At the finer micro-evolutionary scale, all analyses consistently grouped the samples into four geographic units comprised of (1) LOD, KOG, TRY, and SUS; (2) KRO, VID, and GUD; (3) OND; and (4) POL. Anadromous ide populations are also characterized by lower levels of genetic differentiation than freshwater resident ones. This observation is in accordance with previous published comparisons of genetic differentiation between anadromous and freshwater species (e.g., Atlantic salmon *Salmo salar* (Tonteri et al. 2007) and species with both brackish water and freshwater resident populations (e.g., zander *Sander lucioperca* (Säisä et al. 2010). The level of genetic diversity within the anadromous group is also similar to that of other species (perch *Perca fluviatilis* and whitefish *Coregonus maraena*) with possible interpopulation gene flow in the Baltic Sea (Olsson et al. 2011, 2012). Given this, we hypothesize that one factor affecting the observed genomic population structure of ide is gene flow between anadromous populations in regions with salinities within the tolerance of ide, and reduction or cessation of gene flow in regions with salinities above the ide’s salinity tolerance. This hypothesis is further supported by observations of ide in marine habitats occurring in these regions (Fig. 1, 2).

Although the salinity levels allow for frequent gene flow among the four anadromous sampling localities, our results suggest some degree of population genomic structure among these sites. In line with previous observations of freshwater resident ide populations (Winter and Freidrich 2003), as well as anadromous fish species in general (McDowall 2001), we suggest that this fine-scale genetic structuring derives from an affinity of individuals to spawn at their natal site (i.e., homing). Our observation that the geographically close TRY and KOG streams form a single population may result from a lack of analytical power to differentiate them, or from the fishes’ inability to distinguish between geographically close streams, as has been reported for anadromous species such as the alewife *Alosa pseudoharengus* and blueback herring *Alosa aestivalis* (Palkovacs et al. 2014).

In the future, when ocean water levels rise and the salinities change in river deltas around the world, anadromous populations will be facing new challenges. For well-studied, salinity tolerant, anadromous species the focus has been on temperature (Reist et al. 2006; Jonsson and Jonsson 2009; Hedger et al. 2013), but for anadromous populations of true freshwater fishes that exhibit a wider
temperature range, increased salinity could play a more important role. For the populations of anadromous ide that are already close to their maximum salinity tolerance, our data indicate that potential increases in salinity will almost certainly drive populations to isolation, as seen in the current freshwater populations. As a result, we predict increased genetic differentiation between populations, decreased genetic diversity, and the restriction of anadromous behavior to regions in the eastern Baltic Sea with lower and more stable salinity levels. In contrast, should salinity in the western Baltic Sea decrease, the effects will be increased gene flow, which could lead to loss of local adaptive traits, or alternatively, the genetic “rescue” of isolated populations.

Local genomic population structure and possible effects of human impact?

In the cluster that includes all the anadromous populations, we note that the TRY and KOG populations are more closely connected to the Swedish sample site LOD, located 55 km away on the other side of the high-current Øresund strait, than they are to the SUS sample site located 120 km along the coast. If this is a result of modern events it could suggest that ide migrate not only along the coast-line, but also cross deeper waters, and indicate that the waterway distance might play an important role in limiting gene flow. However, this cannot be determined in this study, thus future telemetric studies of ide in this region would represent a means to shed light on such behavior. The freshwater resident population POL is closely related to the anadromous sample sites TRY, KOG, LOD, and SUS. This close relationship could be explained by joined postglacial history or maybe POL could historically have been anadromous with low levels of gene flow, as Lake Arresø was a fjord until the 17th century (Naturstyrelsen 2015).

Sample sites GUD, KRO, and VID are all located on the European mainland, and contain ide belonging to a single genetic group. Among these, the similarity of the KRO and GUD sample sites is particularly noteworthy. The population present at the GUD site was first noticed in the 1970’s, and its gradual movement downstream in the GUD system has since been reported by both local anglers and monitoring programs (Jensen 2006; Carl 2012a). Although these ide are believed to have originated from a single stocking event of 400–500 individuals, the origins of these fish are unknown. In this regard, it is possible that they were illegally imported from Germany, something that has been reported in this region for at least one species of reptile (Jensen 2002) and Wels catfish (*Silurus glanis*) (Carl 2012b). The KRO sample site is part of the Rhine water system that runs through Germany. Furthermore, a study from another central European river showed that ide living in a 120 km stretch is a “single panmictic unit” (Wolter et al. 2003), supporting the hypothesis that German ide might be closely related to ide from the KRO sample site in the Netherlands. Further support for a hypothesis of ide introduction from the Rhine into GUD comes from the observation that \( F_{ST} \) value of the comparison between GUD and KRO ide is lower than that for other landlocked sample sites. On the other hand, one would expect a lower level of heterozygosity as a consequence of a recent founder effect (Nei et al. 1975; Wright 1990), but we do not see this for GUD ide. However, if the GUD ide consist of a mix of original and introduced specimens, genetic diversity may be higher than expected. Therefore, the true origin of this population will need to be further studied before any firm conclusions can be made.

A last point of note relating to the ide in this study, are those from the Odense A (OND) stream. Archaeological excavations have indicated that this population has been present since at least the Iron Age (Gotfredsen et al. 2009), and until the 1930’s they were so abundant that they were used for fertilizing the local fields (Frederiksen 1979). However, during industrialization of the region between the 1940’s and 1980’s the river was used as an open sewer, leading to dramatic declines in the ide population (Carl 2012a). Given this, we note that our observation of a genetically unique population there today, exhibiting variation at levels similar to other sample sites, suggests that the population nevertheless not only survived, but may have avoided a genetic bottleneck despite this pollution (Bickham et al. 2000).

In summary, our results provide a first look into the genomic population structure of ide in north-western Europe, and lay the foundation for further studies. In this regard, telemetric studies of migration in anadromous populations, and the search for genetic markers linked to salinity adaptation, will be of considerable interest. Another research field that could build upon the present study is analyses of postglacial colonization and demographic history. The ide is a species well suited for such studies as it has not been systematically stocked and redistributed by man. During the last decade we have seen an increased interest in the ide from recreational anglers practicing catch-and-release. This may lead to a change in status of the ide from nuisance fish and agricultural fertilizer, to an appreciated recreational species and maybe even future introduction of management strategies.

Acknowledgments

The authors would like to thank Marcus Anders Krag, Nicolai Ulrik Hansen, Lasse Birch Hørup, Jeppe Birch Hørup, Emil Flindt Christensen, Berend Masselink, Ulrik...
Nielsen, Toke Wistoft, and Calle Larsson for help with sampling ide and sharing information on their local waters and Thomas Pålson for help with figure layout. Thanks to BRC Genomic Diversity Facility, Institute of Biotechnology, Cornell University, Ithaca, NY USA for providing a great service. This study was funded partly by Aage V. Jensens Nature Foundation and a postdoctoral grant from the Villum Foundation to Olsen MT, and a DFF-MOBILEX grant (DFF- 1325-00136) to Vieira FG.

**Conflict of Interest**

None declared.

**References**


population (Esox lucius L.), assessed using microsatellite DNA analysis of historical and contemporary samples. Heredity 95:136–143.


Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Figure S1.** Mean DNA concentrations, mean reads, mean tags and mean reads/tag for all sample sites.

**Figure S2.** Isolation by distance in ide in north-western Europe.

**Figure S3.** Cross Validation (CV) error rate of admixture analysis of ide.

**Table S1.** Geographic distance (km) between ide *Leuciscus idus* sample sites. Euclidean distance (top) and waterway distance (below).
Appendix II

Hologenomic Adaptations Underlying the Evolution of Sanguivory in the Common Vampire Bat
Hologenomic adaptations underlying the evolution of sanguivory in the common vampire bat

M. Lisandra Zepeda Mendoza1,19*, Zijun Xiong2,3,19, Marina Escalera-Zamudio4, Anne Kathrine Runge5, Julien Thézé6, Daniel Streicker6, Hannah K. Frank7, Elizabeth Loza-Rubio8, Shengmao Liu3, Oliver A. Ryder9, Jose Alfredo Samaniego Castruita1, Aris Katzourakis5, George Pacheco1, Blanca Taboada10, Ulrike Löber1, Oliver G. Pybus5, Yang Li13, Edith Rojas-Anaya9, Kristine Bohmann1, Aldo Carmona Baez1,11, Carlos F. Arias10, Shiping Liu3, Alex D. Greenwood4,12, Mads F. Bertelsen13, Nicole E. White14,15, Michael Bunce14,15, Guojie Zhang2,3,16, Thomas Sicheritz-Pontén17* and M. P. Thomas Gilbert1,15,18*

Adaptation to specialized diets often requires modifications at both genomic and microbiome levels. We applied a hologenomic approach to the common vampire bat (Desmodus rotundus), one of the only three obligate blood-feeding (sanguivorous) mammals, to study the evolution of its complex dietary adaptation. Specifically, we assembled its high-quality reference genome (scaffold N50 = 26.9 Mb, contig N50 = 36.6 kb) and gut metagenome, and compared them against those of insectivorous, frugivorous and carnivorous bats. Our analyses showed a particular common vampire bat genomic landscape regarding integrated viral elements, a dietary and phylogenetic influence on gut microbiome taxonomic and functional profiles, and that both genetic elements harbour key traits related to the nutritional (for example, vitamin and lipid shortage) and non-nutritional (for example, nitrogen waste and osmotic homeostasis) challenges of sanguivory. These findings highlight the value of a holistic study of both the host and its microbiota when attempting to decipher adaptations underlying radical dietary lifestyles.

The order Chiroptera (bats) exhibits a wide variety of dietary specializations, and includes the only three obligate blood-feeding mammalian species, the vampire bats (family Phyllostomidae, subfamily Desmodontinae). Blood is a challenging dietary source because it consists of an ~78% liquid phase and a dry-matter phase consisting of ~93% proteins and only ~1% carbohydrates, providing very low levels of vitamins, and potentially containing blood-borne pathogens. Vampire bats have evolved numerous key physiological adaptations to this lifestyle, for which the associated genomic changes have not yet been fully characterized due to the lack of an available reference genome. These adaptations include morphological specializations (such as claw-thumbed wings and craniofacial changes including sharp incisors and canines), infrared sensing capacity for the identification of easily accessible blood vessels in prey, and renal adaptations to the high protein content in its diet (such as a high glomerular filtration rate and effective urea excretion). Furthermore, given the high risk of exposure to blood-borne pathogens, another important trait in the vampire bat is its immune system.

Besides genomic adaptations, host-associated microbiota may play an additional, possibly equally important, role in the evolution of vertebrate dietary specialization. Although the functional role of the vampire bat gut microbiome has not been studied, analyses of obligate invertebrate sanguivorous organisms have shown that the gut microbiota contributes to blood meal digestion, provision of nutrients absent from blood and to immunological protection. Studies on mammals have shown that the gut microbiome is a key aspect of an organism’s digestive capacities (energy harvest, nutrient acquisition and intestinal homeostasis), and that it also affects phenotypes related to the immune and neuroendocrine systems. Furthermore, changes in the gut microbiome are associated with diseases such as diabetes, obesity, irritable bowel syndrome and...
inflammatory bowel disease\textsuperscript{15–17}. In response to the growing awareness of the key roles that host–microbiome relationships can play across the spectra of life, various studies have advocated for the ‘hologenome’ concept\textsuperscript{18–20}. In brief, this argues that natural selection acts on both the host and its microbiome (together forming the holobiont); thus, evolutionary studies should incorporate both. The extreme dietary adaptation of vampire bats provides a suitable model to investigate the effect of selection across the genome and microbiome, and thus allows exploration of the role of host-associated microbiome in the evolution of specialized diets.

Results and discussion

Here, we explore the contributions of both the common vampire bat’s nuclear genome and gut microbiome to its adaptation to obligate sanguivory. To this end, we generated both the common vampire bat genome and fecal metagenomic data sets as a proxy to study its gut microbial as well as fecal metagenomic data sets from other non-sanguivorous bat species. We used these data sets for comparative genomic and metagenomic analyses. Specifically, we analysed the common vampire bat genomic landscape, the ratio of substitution rates at non-synonymous and synonymous sites (dN/dS), putative gene loss and gene family expansion/contraction, and computational predictions of the functional impact of amino-acid substitutions. We also performed microbial taxonomic and functional profiling, identified the microbial taxonomic and functional core of the common vampire bat, and identified enriched microbial taxa and functions. Following a holobiongic approach, we identified elements in both the host genome and microbiome that could have played relevant roles in adaptation to sanguivory.

Genomic landscape. We sequenced and de novo assembled the ~2 gigabase (Gb) common vampire bat genome using Illumina sequencing technology (Supplementary Information 1). The genome is smaller than that of other mammals, but similar to previously reported bat genomes\textsuperscript{21}. The initial assembly (~100× mean coverage, scaffold N50 = 5.5 Mb and N90 = 933 kb; Supplementary Figs. 1 and 2) was subsequently improved using the in vitro proximity ligation-based technology for assembly contiguity refinement developed by Dovetail Genomics\textsuperscript{22}. We obtained a final high-quality assembly with scaffold N50 = 26.9 Mb and N90 = 9.46 Mb, contig N50 = 36.6 kb and N90 = 8.8 kb (Supplementary Table 1, Supplementary Fig. 3 and Supplementary Information 1). We used our annotated common vampire bat genome (see Methods—Protein-coding gene and functional annotation) for comparative genomic analyses with publicly available bat genomes and other mammalian genomes (Supplementary Table 2 and Fig. 1).

Repetitive elements can significantly contribute to genome evolution. Thus, for a genomic landscape characterization, we first compared transposable elements in the common vampire bat genome to those within the genomes of non-phyllostomid bats with other diets: the carnivorous Megaderma lyra (greater false vampire bat, Megadermatidae), the insectivorous Pteronotus pteronotus (Parnell’s moustached bat, Mormoopidae) and the frugivorous Pteropus vampyrus (large flying fox, Pteropodidae) (Fig. 1, Supplementary Table 3 and Supplementary Note 1). We identified a 1.6- to 2.26-fold higher copy number of the MULE-MuDR transposon in the common vampire bat genome relative to the other bat genomes. The high mutagenic capacity of MULE-MuDR has been demonstrated to have played a critical role in the evolution of some plants\textsuperscript{23}. Furthermore, transposable elements in general may cause structural or functional changes within the genome and alter epigenetic regulation of the genes into which they are inserted\textsuperscript{24}. Therefore, we explored whether these elements might have also played a role in the evolution of sanguivory by analysing their genomic location in the nuclear genome of the common vampire bat. We found that the identified common vampire bat transposable elements, MULE-MuDR elements in particular, were located within genomic regions enriched for gene ontology (GO) functions related to the challenges of sanguivory, such as antigen processing and presentation, defence response to viruses, lipid metabolism, and vitamin metabolism (Supplementary Information 2).

A sanguivorous diet facilitates exposure to blood-borne viruses that could lead to an increase in genomic invasion by retroviral and non-retroviral endogenous viral elements (EVEs). Thus, we next characterized their presence in the common vampire bat genome. Compared to previously published EVE studies on non-Chiropteran mammals, the common vampire bat exhibits a greater diversity of non-retroviral EVEs in terms of the number of integrations, including endogenized viral genes from avian Bornaviridae.
and Parovoviridae/Dependovirus. However, these findings are not restricted to vampire bats and are similar to those in other bat species15 (Supplementary Note 2, Supplementary Fig. 4 and Supplementary Information 3). Surprisingly, and in contrast to the prior expectations given its sanguivorous diet, the diversity of endogenous retroviral elements (ERVs) in the common vampire bat is very low compared to other bat species16. The only proviral elements detected were DrERV and DgERV, both present in low copy numbers (Supplementary Note 3, Supplementary Fig. 5 and Supplementary Information 4). We hypothesize that genome colonization by ERVs could have been restricted by the genomic adaptations in the common vampire bat genome against ERV insertion and proliferation. In support of our hypothesis, we identified expansion of the anti-retroviral gene TRIM5 family (Viterbi P = 0.000088, Supplementary Information 5 and Supplementary Note 4).

Genomic adaptations to sanguivory. Feeding specialization often requires morphological and physiological adaptations in traits such as the sensory apparatus (for example, infrared sensing), locomotion, digestion, kidney function and immunity (Fig. 2a and Supplementary Information 6). For example, it has been shown that vampire bats have a loss of sweet taste genes and reduction of bitter taste genes17. In agreement, such genes were also identified in our putative gene loss analysis (Supplementary Information 7 and Supplementary Note 5). It is likely that the function of those genes is related to sanguivory, because sweet and bitter taste receptor genes influence glucose homeostasis in humans18. Interestingly, we found that the common vampire bat bitter taste receptor gene TASSR3 has experienced episodic positive selection and shows two species-specific positively selected sites (PSSs) on topological domains, one of them having a potential impact on protein function (PROVEAN score = −4.4) (Supplementary Table 4 and Supplementary Fig. 6). Among the enriched GOs of the differentially evolving genes, we identified functions related to the regulation of RNA splicing, which could be relevant to sanguivory given that D. rotundus produces submandibular tissue-specific splicing isoforms to counteract the prey's response to injury19 (Supplementary Information 8 and Supplementary Note 6). Regarding the recruitment of alternatively spliced forms, the ganglion-specific splicing of TRPV1 has been found to underlie the vampire bat's infrared sensing ability20. Interestingly, we found that PKD1, which directly modulates the product of TRPV1, is positively selected and exhibits species-specific PSSs (branch-site test P = 3.39 × 10−10, branch test P = 1.39 × 10−6, Supplementary Information 8). These examples suggest an important role of alternative splicing as a form of regulatory evolution fundamental to sanguivory (Supplementary Note 7). However, it is clear that despite the number of detected genomic features related to sanguivory adaptation, they alone cannot address all of the challenges posed by this diet (Fig. 2b).

Gut microbiome diet and phylogenetic influence. We generated Illumina shotgun metagenome data in order to compare the gut microbiomes of 13 faecal samples from common vampire bats with those of non-sanguivorous non-phyllostomid bats; specifically, eight frugivorous Rousettus aegyptiacus (Egyptian fruit bat, Pteropodidae), five insectivorous Rhinolophus ferrumequinum (greater horseshoe bat, Rhinolophidae) and five carnivorous Macroderma gigas (ghost bat, Megadermatidae) bats (Supplementary Information 9). We obtained a median of 15.8 Gb of sequencing data (~37.6 million 100-bp paired-end reads) for each dietary category. After filtering low-quality bases, adapter sequences and bat-genome-derived reads, we obtained a median of 2.77 Gb of high-quality data for each species, totalling 86.73 Gb of data. We identified taxa and functions present only in the common vampire bat microbiome (gut microbiome core), as well as taxa and functions that exhibit statistically significant differences in abundance or contribution to variation between the different microbiomes (Supplementary Information 6 and 8, Supplementary Tables 5 and 6 and Supplementary Note 8).

It has been observed previously that similarity in the taxonomic composition of vertebrate gut microbiomes (including bats) can be influenced by the diet and the phylogenetic relationships of the respective host species21. Overall, the common vampire bat microbiome composition is more similar to that of the insectivorous and carnivorous bats than to that of the frugivorous bat. This may reflect a phylogenetic influence on the microbiome taxonomic profile (Fig. 3a, Supplementary Fig. 7 and Supplementary Notes 9 and 10). In contrast, the vampire bat microbiome is strikingly different to that of the compared bats at the functional level, which was characterized by the KEGG annotations of the microbial non-redundant gene catalogues assembled from the metagenomic data sets. While there is little differentiation between the functional gut microbiomes of carnivorous, insectivores and frugivorous bats, the common vampire bat functional gut microbiome is almost completely distinct, and exhibits the least intra-species variation between the samples (Fig. 3b, Supplementary Fig. 8 and Supplementary Table 7).
This suggests that the functional profile is less influenced by phylogeny than the taxonomic profile, and that the common vampire bat gut microbiome harbours a set of functions specialized to its extreme diet (Supplementary Note 11). Subsequently, we analysed the comparative genomic and metagenomic results in a hologenomic framework to demonstrate how both components contribute to adaptation to sanguivory (Fig. 4, Supplementary Figs. 9–12, Supplementary Tables 7–9 and Supplementary Information 8 and 10).

The hologenomic framework of sanguivory. Viscosity and subsequent coagulation represent a challenge for ingestion and digestion of blood. Besides developing potent anticoagulants in its saliva, the common vampire bat holobiont addresses this challenge in various ways. For instance, REG4, involved in metaplastic responses of the gastrointestinal epithelium, was found to be under ongoing positive selection (M8a/M8 test \( P = 0.047 \)) with possible functional impact on its carbohydrate-binding capacity, including binding of the anticoagulant heparin (Supplementary Table 4 and Supplementary Information 8). Furthermore, we identified genes in the common vampire bat microbial functional core from pathways for degradation of heparan sulfate and dermatan sulfate, both being polysaccharides involved in blood coagulation (Supplementary Information 10). We also identified an enrichment in the common vampire bat microbial gene \( t \)-asparaginase (Fishier’s \( P = 0.00027 \)), which decreases protein synthesis of coagulation factors (Supplementary Information 11).

Besides specialized digestion, sanguivory poses other challenges related to the poor nutritional value of blood itself, as well as to the side effects that may arise due to the blood components being the sole dietary source (Fig. 2b). We identified elements in both the genome and gut microbiome that might be involved in solving each of these challenges discussed next.

Hologenomic solutions to nutritional challenges. Low nutrient availability. Obligate sanguivory requires adaptation to very low levels of some nutrients, such as essential amino acids and the vitamin B complex, and very high levels of others, such as salt. Our data clearly demonstrate how both the host and its associated gut microbiome have dealt with these challenges. We found the gene LAMTOR5 to be positively selected in the common vampire bat genome (false discovery rate (FDR) = 2.02 \( \times 10^{-7} \), Supplementary Information 8). This gene is involved in the response to nutrient starvation, suggesting that the common vampire bat metabolism has adapted to the low nutrient content available in blood. Similarly, we identified KOs in the common vampire bat microbial core related to energy and carbohydrate metabolism (Fig. 4b and Supplementary Information 11). For example, when compared to the other bats, we identified an enrichment in the common vampire bat microbial genome (false discovery rate (FDR) = 2.02 \( \times 10^{-7} \), Supplementary Information 8). Furthermore, we found the gene PDZD11 in the common vampire bat genome, involved in vitamin B metabolism, evolved faster in the common vampire bat genome relative to the other examined bats (branch test \( P = 1.97 \times 10^{-10} \)). We further postulate that the microbiome also contributes in tackling the low nutritional challenge by providing necessary nutrients. For example, compared to other bats, the common vampire bat gut microbiome had the highest number of enriched enzymes related to the biosynthesis of cofactors and vitamins, such as carotenoid (Supplementary Fig. 8B and Supplementary Information 10 and 11). Furthermore, we identified enzymes involved in the metabolism of butyrate, an important energy source for the gut microbiome. These findings suggest that the coevolution between host and microbiome has enabled the common vampire bat to successfully adapt to a diet consisting predominantly of blood.
important nutrient for cells lining the mammalian colon derived from bacterial fermentation, enriched in the common vampire bat gut microbiome as well as in the vampire bat gut microbiome core (Supplementary Information 10 and 11).

**Lipid and glucose assimilation.** Besides vitamins and other nutrients, lipids and glucose may not be readily available in blood. Furthermore, vampire bats have a reduced capacity to store energy reserves. In agreement with this, we identified GO enrichment of lipid metabolism on genes with dN/dS values statistically higher or lower compared to other bats (Fig. 4a,c and Supplementary Information 8). For example, we identified the gene FFAR1, which plays an important role in glucose homeostasis, as evolving faster in the common vampire bat compared to the other bats (branch test \( P = 3.68 \times 10^{-4} \)) and containing amino-acid substitutions with a possible functional impact on its binding ability (Supplementary Table 4). This may enable the common vampire bat to better utilize the available glucose. The common vampire gut microbiome also exhibits unique solutions to the challenge (Fig. 4b,d and Supplementary Information 10 and 11). Differences in the carbohydrate and glycans metabolism functional profile were identified in the principal component analysis (PCA) comparing the different microbiomes (Supplementary Fig. 11), which place the vampire bat gut microbiome profile within a cluster separate from those of the non-sanguivorous bats. Importantly, we identified enrichment in the microbial gene glyceraldehyde kinase in the common vampire bat (Fisher's \( P = 0.0027 \)), which plays a key role in the formation of triacylglycerol and in fat storage, and its deficiency causes symptoms such as hypoglycaemia and lethargy in a mouse model.

**Hologenomic solutions to non-nutritional challenges. Immunity.** Due to its sanguivorous lifestyle, the common vampire bat risks direct contact with blood-borne pathogens from prey. Consequently, we observed \( > 280 \) bacterial species known to be pathogenic to some mammalian species present exclusively in the common vampire bat gut microbiome (Supplementary Information 12). For example, we identified enrichment of genes from *Borrelia* (Supplementary Information 11) and *Bartonella* as one of the most abundant genera in the common vampire bat compared to the other bats (Supplementary Information 13). These bacteria are known to be transmitted by sanguivorous invertebrates (ticks, fleas, mosquitoes and lice). This suggests that the abundance of this genus could be a shared pattern of sanguivorous species. While several studies have elucidated part of the expected genomic immunity-related adaptations, analysis of the full genome enabled us to identify more elements related to immunity, such as defence response to virus and antigen processing and presentation (Fig. 4a,c and Supplementary Information 8). For example, we identified the antimicrobial gene *RNASE7* to be positively selected (branch-site test \( P = 0.004 \)) and containing amino-acid substitutions that may increase its bactericidal capacity (Supplementary Table 4). In addition, when compared to the gut microbiomes of non-sanguivorous bats, that of the common vampire bat contains a large abundance of potentially protective bacteria, such as *Amycolatopsis mediterranei* (\( P < 0.05 \)).
which has been shown to produce antiviral compounds against bac-
teriophages and poxviruses\(^5\) (Supplementary Information 13).

**Iron assimilation.** Iron concentration represents a significant chal-
lenge to sanguivory. Although the concentration of free iron is not
high in the blood, severe haemolysis (for example, during diges-
tion of blood) could result in high levels of iron that, if absorbed in
excess, could accumulate and disrupt the normal function in organs
such as the liver, heart and pancreas. Interestingly, we identified the
light and heavy chains of the iron-storing protein ferritin (encoded
by FTL, FTH1) under gene family expansion in the common
vampire bat genome (Viterbi P=0.0012, respec-
tively, Supplementary Information 5). In addition, we identified an
enrichment of the iron-storing protein ferritin (Fisher’s P=0.0014),
suggesting that the gut microbiome also contributes to solving this
challenge (Fig. 4d and Supplementary Information 11).

**Nitrogen waste and blood/osmotic pressure.** The high abundance of
protein in the blood and its rapid ingestion could lead to accu-
mulation of nitrogenous waste products, primarily urea, which
could lead to renal disease-like symptoms (for example, high blood
pressure and fluid retention). This challenge is exacerbated by the
abundance of salts in blood, which pose additional osmotic and
blood pressure challenges. We see at the genome level that this is
addressed by a higher rate of evolution in the common vampire
bat genes compared to the other bats involved in disposal of excess
nitrogen (Fig. 4a,c and Supplementary Information 8), such as
PSMA3 (branch test P=2.08×10\(^{-7}\)). This challenge seems to also
be addressed by the gut microbiome. The PCA of the copy num-
ber of genes involved in amino-acid metabolism distinguishes the
common vampire bat in a single cluster separated from the other
bat species analysed (Supplementary Fig. 11), suggesting a special-
ized microbial amino-acid metabolism capacity. We also identified
enrichment in the common vampire bat microbial gene urease sub-
unit alpha (ureA, Fisher’s P=0.016) involved in urea degradation
(Supplementary Information 11).

**Conclusions**

It is clear from our results that the common vampire bat has adapted
to sanguivory through a close relationship between its genome and
gut microbiome. We identified a phylogenetic and dietary impact
on the common vampire bat gut microbiome and uncovered an
unexpected genomic viral and repetitive element genomic make-
up. We showed that extreme dietary specializations, such as that of
the common vampire bat, provide a comparative framework with
which to tease apart the relative roles of genomes and microbiomes
in adaptation. In conclusion, our study illustrates the benefits of
studying the evolution of complex adaptations under a holobionte
framework, and suggests that vertebrate adaptation studies that do
not account for the action of the holobionte may fail to recover the
full complexity of adaptation.

**Methods**

**Genome sequencing and raw-read processing.** We shotgun-sequenced the D.
**rotundus** genome using a wing biopsy from a sample collected by the NIH through
the Catoctin Wildlife and Zoo in Thurmont, Maryland, USA. The capturing
method and dead preservation procedure of the specimen are unknown. The age
and sex of the dead individual are unknown. Samples permitted were to BGI
for the sequencing of the specimen, originally as part of the BGI 10K genome
project. Genomic DNA was extracted at the Laboratory of Genomic Diversity
using KmerFerov\(^8\) by dividing the total number of seven decamers by the peak of the seven
decamer Poisson distribution. High-quality reads were assembled using SOAPdenovo\(^9\)
follows. (1) Short-insert library reads were assembled as initial contigs ignoring
the sequence pair information. (2) Reads were aligned to the previously generated
contig sequences. Scaffolds were constructed from short-insert-size libraries to
large-insert-size libraries by weighting the paired-end relationships between pairs
of contigs, with at least three read pairs required forming a connection between
any two contigs. (3) Gaps in the scaffolds were closed using CapCloser\(^10\). Genome
quality assessment was performed by downloading a publicly available D.**rotundus**
transcriptome\(^11\) and aligning the transcripts to the genome using BLAT\(^12\).

**Genome contiguity improvement.** We prepared two Chicago libraries\(^13\) using 5 µg
of high-molecular-weight DNA obtained from D. rotundus cultured cells from the San
Diego Zoo collection, which were originally derived from a skin sample taken from
between the shoulder blades of a D. rotundus individual. Permits for this
were obtained from the San Diego Zoo Global. The capturing method and
dead preservation procedure of the specimen are unknown. The age and sex of the
dead individual are unknown. DNA was extracted with QiaQuick Blood and
Cell Midi kits according to the manufacturer's instructions. The steps required for
building the Chicago libraries were performed as described in ref. 17. The libraries
were sequenced using Illumina HiSeq 2500 2×100 bp rapid run. Our initial D.**rotundus**
assembly, shotgun sequence data and Chicago libraries were used by Dovetail Genomics as input data for HiRise, as described in ref. 17. Genome assembly contiguity statistics were obtained using a minimum N track length of 1
to delimit the contig blocks within the scaffolds.

**Protein-coding gene and functional annotation.** Homology-based gene prediction
was performed using as a reference the Ensemble gene sets of Myotis
lucifugus, Pteropus alecto, Myotis davidi, horse and python. We used protein
sequences of the reference gene sets to the D. rotundus assembly using blastx\(^18\)
and linked the blast hits into candidate gene loci with geneBLAST\(^19\). We
filtered out those candidate loci with a homologous block length <90% of the query
length. We extracted genomic sequences of candidate gene loci, including the
intronic regions and 3 kb upstream/downstream sequences. The sequences were
passed to GeneWise\(^20\) to search for accurately-spliced alignments. We filtered out pseudogenes containing more than one frame error for single-exon genes. Potentially pseudogenized single exons were removed if they were part of a multi-
exon gene. We then aligned protein sequences of these genes against UniProt using
BLAST and filtered out genes without matches. We also filtered out genes that had >80%
repeat regions. De novo gene prediction was performed with AUGUSTUS\(^21\) using a published common vampire bat transcriptome\(^22\) as a training data set and with masked transposable-element-related repeats. We filtered out partial and <150 bp
predicted genes. Genes that aligned over 50% of their length to annotated
transposable elements were filtered out. Finally, we built a non-redundant gene set
with the homology-based evidence prioritized over the de novo evidence. If the de
novo genes were chosen in the reference gene set, we retained only those
candidates with over 90% of their length aligning against UniProt\(^23\) and that contained at least
3 exons. The integrated gene set was translated into amino-acid sequences,
which were used to search the InterPro database with sprescan, 4.8\(^24\). We used
BLAST to search the metabolic pathway database in KEGG\(^25\) and homologies in the
SwissProt and TREMBL databases in UniProt. The quality and annotation of the D.**rotundus** gene
was assessed and compared to that of D. communis, D. cervinatus, D. vulpinus, M. brandti, M. davidi, M. lucifugus and P. parnellii with BUSCO.

**Repeat annotation.** Repeat annotation was performed on the genomes of D. rotundus, P. parnellii, M. lyra and P. parnellii. The motifs
were identified using RepeatMasker\(^26\) and RepeatProteinMask against the Repbase transposable element library\(^27\). We used RepeatCount, PILER-DF and
RepeatModeler to construct a de novo transposable element library, which was then used by RepeatMasker to predict repeats. We predicted tandem repeats using TRF\(^28\). LTRFinder\(^29\) was used to detect long terminal repeats (LTRs). The Repbase-based annotations and the de novo annotations were merged.

**Non-retroviral EVEs.** We constructed a comprehensive library of all non-
retroviral virus protein sequences available in GenBank and EMBL. We used DIAMOND\(^30\) to search these sequences against the D. rotundus
and Myotis lyra genomes. We then extracted the matching amino-acid sequences and performed reciprocal blast-like searches\(^31\) using DIAMOND with the selected subset of D. rotundus amino-acid sequences and the set of non-redundant protein sequences. D. rotundus genomic sequences were considered of viral origin if they unambiguously matched viral
Functional characterization. We performed GO analysis using GOrilla\(^{77}\) as well as manual characterization using the GO annotations of the human genes downloaded from Ensembl\(^{78}\), which was used for a comprehensive characterisation of the functional impact of the non-syonymous substitutions present only on the proteins in \(D.\ rotundus\).

Species-specific PSS identification and protein modelling. We tested for positive selection and positive selected sites (PSSs) using the complete CDS alignments for the proteins FFAR1, PLXNA4, RIGA, RNAs7 and TA2R3 from various species downloaded from the OrthoMCL database\(^{79}\). Results were re-analysed using MUSCLE\(^{80}\) in SeaView\(^{81}\) and phylogenetic analysis was performed using PhyML\(^{82}\). We used PAML codeml\(^{83}\) to test the M1/M2 branch model constraining the \(D.\ rotundus\) node\(^{84}\), the M8a/M8 site model, and the branch-site model A constraining the \(D.\ rotundus\) node. The branch-site model A was evaluated under a LRT against the null hypothesis, while PSSs were scored under naive empirical Bayes (NEB) and available at the bat gene syndrome (NEB) and available at the bat gene syndrome.

Metagenomic data. We used fecal and anal swab samples from \(D.\ rotundus\), \(R.\ aegyptiacus\) and \(M.\ parnellii\). The \(D.\ rotundus\) sequences were obtained from wild \(D.\ rotundus\) individuals captured using mist nets and cloth bags and preserved in RNA later according to the manufacturer’s instructions. We used Illumina 2500 HiSeq platform. For each putative viral peptide, we retrieved the function and predicted the taxonomic assignment by comparison to the best reciprocal blastp-like hit protein. Phylogenetic analyses were performed by aligning the sequences using MAFFT\(^{63}\) and CDS alignments from each data set were obtained using an in-house script with the MGmapper tool. We then performed de novo assembly using the MGmapper. Using these results, we performed de novo assembly using the MGmapper. Using these results, we performed de novo assembly using the MGmapper.

Orthologous gene families. Using the \(D.\ rotundus\) genome against a range of other mammalian species, we performed clustering of orthologous genes using two strategies. (1) Identifying single-copy orthologs in the species by using the reciprocal best hit method\(^{85}\). The homeologous set in the Refseq database from NCBI and we obtained a non-redundant protein catalogue to be used as a reference by collapsing redundant genes and keeping the longest open reading frame. We used Bos taurus, Equus caballus and \(E.\ s.\ \)europaeus as outgroups. Genes identified as missing in \(D.\ rotundus\) were searched with blastp against the protein catalogue of the other available databases. Further validation on selected genes was performed by evaluating the conservation of their syntenic regions compared to other species, and the GC context of the syntenic regions compared to the \(D.\ rotundus\) genome average GC content, searching them against the published \(D.\ rotundus\) transcriptome, and through PCR assays.

Putative gene loss. We identified genes putatively lost in \(D.\ rotundus\) as previously described in ref.\(^{86}\). The homeologous set in the Refseq database from NCBI and we obtained a non-redundant protein catalogue to be used as a reference by collapsing redundant genes and keeping the longest open reading frame. We used Bos taurus, Equus caballus and \(E.\ s.\ \)europaeus as outgroups. Genes identified as missing in \(D.\ rotundus\) were searched with blastp against the protein catalogue of the other available databases. Further validation on selected genes was performed by evaluating the conservation of their syntenic regions compared to other species, and the GC context of the syntenic regions compared to the \(D.\ rotundus\) genome average GC content, searching them against the published \(D.\ rotundus\) transcriptome, and through PCR assays.

Taxonomic and functional metagenomics comparison. We filtered on the basis of the breadth of coverage and the identifications of the extraction blanks. We removed any non-microbial hit and any taxon in which the pairwise reads matched different genera or only one of the reads had a hit. The counts were normalised by percentage. We identified a microbial taxonomic and functional sanguivorous core by comparing the filtered sets of the bats and keeping as core those taxa and genes identified only in the \(D.\ rotundus\) samples. We manually examined the taxa from the filtered taxonomic identifications, and the KO and COG annotations from the filtered non-redundant gene set catalogue. We compared the normalized abundance of taxa and functions between \(D.\ rotundus\) and the non-sanguivorous bats as follows. (1) Using the distribution of the different functional categories from \(D.\ rotundus\) and each non-sanguivorous bat species with a Wilcoxon rank-sum test. (2) Using the entire taxonomic and functional data sets, as well as down-sampling the data set filtered from Ensemble to the minimum, median and third-quartile values of the count distributions. With the resulting data sets, we calculated the Euclidean, Bray–Curtis and Jaccard distance metrics with the R package vegdist\(^{87}\), and used the Ward hierarchical clustering method using UPGMA and Ward, and the non-sanguivorous bats as well. (3) We identified taxa and functions significantly contributing to the variation between the \(D.\ rotundus\) and the non-sanguivorous bat species. We examined the rotation matrix from the PCA of the normalized counts, excluding the four deepest sequenced samples, of the species and genus microbial taxonomic levels and the KEGG functional pathways. We identified the most significantly abundant
D. rotundus microbial taxa as those with a significantly higher median normalized count value (FP<0.05) in D. rotundus and a median and mean normalized count value of 0 in the other 3 bat species for the first 3 principal components. We also identified significantly more abundant genes in D. rotundus by generating and annotating with KEGG a non-redundant gene set with all of the predicted genes from all of the bat samples. The reads of the samples were mapped against this bat non-redundant gene set, and a normalized count matrix was generated and used for Fisher tests on each of the functional pathways.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. In house scripts used for the processing of the data are available from the corresponding authors upon request.

Data availability. The NCBI BioProject accession code for the genome assembly is PRJNA414273, and the sequence reads are available at the NCBI sequence read archive (SRA) under the accession SRA619672. The BioProject code for the metagenomic sequencing data is PRJNA115000, and the reads can be accessed at SRA with the accession SRA209977.

Received: 23 August 2016; Accepted: 11 January 2018; Published online: 19 February 2018

References

29. Hayward, J. A. et al. Identification of diverse full-length endogenous viral genomes as those with a significantly higher median normalized count value (FP<0.05) in D. rotundus and a median and mean normalized count value of 0 in the other 3 bat species for the first 3 principal components. We also identified significantly more abundant genes in D. rotundus by generating and annotating with KEGG a non-redundant gene set with all of the predicted genes from all of the bat samples. The reads of the samples were mapped against this bat non-redundant gene set, and a normalized count matrix was generated and used for Fisher tests on each of the functional pathways.

NATURE ECOLOGY & EVOLUTION | www.nature.com/teco

Acknowledgements
We thank M.I. Mendoza-Ojeda, G. Czirjak, L. Caballero, O. Rios, A. Patraca, C. Tello, D. Becker and M.E. Roelke-Parker for the collection and exportation of vampire bat samples, as well as P. Perelman and M. Korody for the preparation of genomic DNA. We thank Dovetel Genomics for the improvement of the genome contiguity of D. rotundus, in particular R. E. Green, M. Hartley and R. Rice. We also thank G. Jones for his support in the capture of the R. ferrumequinum bat samples and for useful comments on the manuscript, and R. Dunn and L. Nichols for their support at the early stages of the project. Likewise, we thank A. Albrecht and O. Azpauru for assistance regarding samples from Rhinolophus euryale, Miniopterus schreibersii, Nyctalus leisleri and Nyctalus lasiopterus. We thank the Danish National High-Throughput DNA Sequencing Centre for assistance in generating the metagomome and Dovetel Illumina sequence data. We also thank the Instituto de Biocenticologia-UNAM for giving us access to its computer cluster and J. Verleyen for his computer support. M.T.G.P. and M.L.Z.M recognize Lundbeck Foundation R52-A5062, ERC Consolidator Grant (683199 – Extinction Genomics) and Danish National Research Foundation DNF194 grants as the principal funding behind this work. M.E.-Z. and A.D.G. acknowledge the Deutsche Forschungsgemeinschaft (DFG) (KR 3924/9-1) to A.D.G. and an international doctoral scholarship provided by the Consejo Nacional de Ciencia y Tecnología (CONACyT) and the German Academic Exchange Service (DAAD) (316495-GOP) and T. acknowledge funding from the European Research Council under FP7/2007-2013/ERC grant agreement no. 614725-PATHPHYLODYN. A.K. recognizes funding from the Royal Society. H.K.F. acknowledges funding from a Ting and Mary MacDonald fellowship in Environmental Science and Conservation. Field work in Peru was funded by the US National Science Foundation (grant DEB-1029966) and a Sir Henry Dale Fellowship to D.S., jointly funded by the Wellcome Trust and the Royal Society (grant 102507/Z/13/Z). Funding for Costa Rican fieldwork came from the Stanford Woods Institute for the Environment Environmental Venture Project.

Author contributions

NATURE ECOLOGY & EVOLUTION | www.nature.com/natcoevol
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
ARTICLES

Competing interests
The authors declare no competing financial interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0476-8.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.L.Z. or T.S. or M.P.T.G.

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

Killer whale genomes reveal a complex history of recurrent admixture and vicariance


*email: footead@gmail.com

1CMPG, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland
MEFGIL
2Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, Gwynedd, LL57 2UW, UK
3NTNU University Museum, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
4Natural History Museum of Denmark, University of Copenhagen, Øster Volgade 5-7, Copenhagen K 1350, Denmark
5Scottish Oceans Institute, East Sands, University of St. Andrews, St. Andrews, Fife KY16 8LB, UK
6Marine Mammal and Turtle Division, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanographic and Atmospheric Administration, 8901 La Jolla Shores Drive, La Jolla, California 92037, USA
7Greenland Institute of Natural Resources, Kivioq 2, PO Box 570, 3900 Nuuk, Greenland
8American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024-5192, USA
10Cascadia Research, 4th Avenue, Olympia, Washington 98501, USA
11Marine Mammal Institute, Department of Fisheries and Wildlife, Oregon State University, Newport, HMSC, 2030 SE Marine Science Drive, Oregon 97365, USA
12School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
13Scottish Marine Animal Stranding Scheme, SRUC Veterinary Services Drummondhill, Stratherrick Road, Inverness, IV2 4JZ, UK
14The Ocean Giants Program, Wildlife Conservation Society, New York, USA
15Observatório Pelagis, UMS 3419, Université de La Rochelle-CNRS, Université de La Rochelle, 5 allées de l’océan, 17000 La Rochelle, France
16Laboratório de Ecologia e Conservação da Megafauna Marinha, Instituto de Oceanografia, Universidade Federal do Rio Grande, Av. Itália km. 8 s/n, Campus Carreiros, Rio Grande, RS 96201-900, Brazil
17CIRCE, Conservation, Information and Research on Cetaceans, Cabeza de Manzaneda 3, Algeciras, 11390, Spain
18Fisheries & Oceans Canada, 501 University Crescent, Winnipeg, Manitoba R3T 2N6, Canada
Abstract

Reconstruction of the demographic and evolutionary history of populations assuming a consensus tree-like relationship can mask more complex scenarios, which are prevalent in nature. An emerging genomic toolset, which has been most comprehensively harnessed in the reconstruction of human evolutionary history, enables molecular ecologists to elucidate complex population histories. Killer whales have limited extrinsic barriers to dispersal and have radiated globally, and are therefore a good candidate model for the application of such tools. Here, we analyse a global dataset of killer whale genomes in a rare attempt to elucidate global population structure in a non-human species. We identify a pattern of genetic homogenisation at lower latitudes and the greatest differentiation at high latitudes, even between currently sympatric lineages. The processes underlying the major axis of structure include high drift at the edge of species’ range, likely associated with founder effects and allelic surfing during post-glacial range expansion. Divergence between Antarctic and non-Antarctic lineages is further driven by ancestry segments with up to four-fold older coalescence time than the genome-wide average; relicts of a previous vicariance during an earlier glacial cycle. Our study further underpins that episodic gene flow is ubiquitous in natural populations, and can occur across great distances and after substantial periods of isolation between populations. Thus, understanding the evolutionary history of a species requires comprehensive geographic sampling and genome-wide data to sample the variation in ancestry within individuals.
1 | INTRODUCTION

Genetic divergence of isolated populations can be interrupted by episodic gene flow during periods of spatial contact, which can erode genetic differences between populations (Durand et al., 2009; Gompert et al., 2010), rescue small isolated populations (Frankham, 2015), and maintain standing genetic variation that can act as a substrate for future adaptation (Brawand et al., 2014; Meier et al., 2017). The geographic context of ancestral spatial contact is difficult to elucidate from the contemporary distribution of modern populations (Pickrell & Reich, 2014; Foote, 2018; Peñalba et al., 2018), especially in marine species, which can have dynamic ranges due to the low energetic cost of movement and few physical barriers in the oceans (Gagnaire et al., 2015; Kelley et al., 2016). Additionally, ancestral episodes of admixture may have occurred via now-extinct ‘archaic’ populations (Racimo et al., 2015) or sister species (Fraïsse et al., 2015), further complicating the inference of the biogeographic history of ancestry components from the spatial distribution of contemporary populations. However, periods of admixture leave genomic signatures that can be used to infer the direction, extent and timing of ancestral gene flow (Patterson et al., 2012; Sousa & Hey, 2013; Racimo et al., 2015; Duranton et al., 2018). Divergence-with-gene-flow is often studied at local scales, but could influence global genetic structure and variation through connected networks of populations (Novembre et al., 2008; Booth Jones et al., 2018).

Killer whales have a global distribution rivalling that of modern humans, yet they can exhibit fine-scale geographic variation in ecology and morphology (Ford et al., 1998; Durban et al., 2017), reflecting variation in their demographic and evolutionary history (Hoelzel et al., 2007; Morin et al., 2015; Foote et al., 2016). The best-studied ecotypes are the mammal-eating ‘transients’ and fish-eating ‘residents’ found in partial sympatry throughout the coastal waters of the North Pacific (Ford et al., 1998; Saulitis et al., 2000; Matkin et al., 2007; Filatova et al., 2015). Four decades of field observations have found that residents and transients are socially isolated and genetically differentiated across their geographic range (Hoelzel & Dover, 1991; Barrett-Lennard, 2000; Ford, Ellis & Balcomb, 2000; Hoelzel et al., 2007; Morin et al., 2010; Parsons et al., 2013; Filatova et al., 2015). There has been a contentious debate regarding whether the formation of these two ecotypes was initiated in sympatry (Moura et al., 2015), or results from secondary contact of two distinct lineages (Foote et al., 2011; Foote & Morin, 2015, 2016).
In the waters around the Antarctic continent, killer whales have diversified into distinct morphs, partially overlapping in their ranges (Pitman & Ensor, 2003; Pitman et al., 2007; Durban et al., 2017). These include type B1, which is commonly observed hunting seals in the pack-ice; type B2, which has been observed foraging in open water for penguins; and type C, which is most commonly observed in the dense pack-ice, and is thought to primarily feed on fish (Pitman & Ensor, 2003; Pitman & Durban, 2010, 2012; Durban et al., 2017). Perhaps surprisingly given their highly distinct morphological forms (Pitman & Ensor, 2003; Pitman et al., 2007; Durban et al., 2017), the Antarctic types are inferred from previous genomic analyses to have diversified from a recent shared ancestral lineage following an extended genetic bottleneck (Morin et al., 2015; Foote et al., 2016). However, the reconstruction of the evolutionary relationships among ecotypes, and how these relate to a more globally distributed dataset, has proved challenging due to incomplete lineage sorting and admixture, and a paucity of genomic data from a wider geographic distribution (Foote & Morin, 2016).

Here we fill this gap by providing 27 additional genomes to a global dataset totalling 47 genomes and trace population history of separation and admixture. First, we describe the global patterns of biodiversity, then we focus on the history of well characterised ecotypes of the North Pacific and Antarctica to test between the opposing hypotheses of simple history of vicariance versus a more complex (and previously hidden) history involving ancient secondary contact between divergent lineages.

2 | MATERIALS AND METHODS

2.1 | Dataset

Genome sequences were generated from 27 individuals that best represented the known global geographic, genetic and morphological diversity of this species (Figure 1a). For a subset of analyses, we further included 20 previously sequenced genomes (Supporting Information Table S1): four additional genomes each from the North Pacific transient and resident ecotypes, and Antarctica types B1, B2 and C (Foote et al., 2016), and 10 RAD-seq generated genotypes (Moura et al., 2015). In addition, we sequenced an outgroup sample of a long-finned pilot whale (Globicephala melas) from a mass stranding at Ratmanoff beach, Kerguelen island in the Southern Ocean and included sequencing reads of the bottlenose dolphin (Tursiops truncatus, Short Read Archive accession code SRX200685).

2.2 | Library building, sequencing and mapping.
Samples were selected from a global dataset of 452 individuals that best represented the known global geographic and genetic diversity of this species (Morin et al., 2015). Where possible, we selected identifiable individuals from longitudinally studied populations, e.g. Crozet Archipelago (Guinet & Tixier, 2011); Gibraltar (Esteban et al., 2016); and Iceland (Samarra & Foote, 2015). DNA was extracted from skin biopsies, with the exception of the recently described type D morphotype (Pitman et al., 2011), from which dry soft tissue and powdered tooth were sampled from a specimen (#1077) which stranded on Paraparaumu Beach, New Zealand in 1955, and is now part of the collections of the Museum of New Zealand, Te Papa Tongarewa, Wellington. This is the only available sample for DNA analysis from this rarely observed morphotype, which has a pelagic circumpolar subantarctic distribution, making it logistically difficult to biopsy sample (Pitman et al., 2011; Foote et al., 2013; Tixier et al., 2016).

DNA was extracted using a variety of common extraction methods as per Morin et al. (2015). Genomic DNA was then sheared to an average size of ~500 bp using a Diagenode Bioruptor Pico sonication device. The sheared DNA extracts were converted to blunt-end Illumina sequencing libraries using New England Biolabs (Ipswich, MA, USA) NEBNext library kit E6040L. Libraries were subsequently index-amplified for 20 cycles using a KAPA HiFi HotStart PCR kit (Kapa Biosystems, Wilmington, Ma. USA) in 50-µl reactions following the manufacturer’s guidelines. The amplified libraries were then purified using a QIAquick PCR purification kit (Qiagen, Hilden, Germany) and size-selected on a 2% agarose gel in the range 422-580 bp using a BluePippin instrument (Sage Science, Beverly, MA, USA). The DNA concentration of the libraries was measured using a 2100 Bioanalyzer (Agilent Technologies, CA, USA); these were then equimolarly pooled and sequenced across four lanes of an Illumina HiSeq4000 platform using paired-read PE150 chemistry and two lanes using single-read SR100 chemistry.

DNA libraries from the soft tissue and powdered bone of the type D museum specimen were initially pooled and shotgun sequenced across one lane of an Illumina HiSeq2000. This indicated a low endogenous content of DNA (1.5% in soft tissue libraries and 4.0% in powdered tooth libraries). Libraries were therefore subjected to two rounds of whole genome enrichment capture using genome-wide biotinylated RNA baits built from modern DNA by Arbor Biosciences, Ann Arbor, MI (Enk et al., 2014), prior to pooling and sequencing across a second lane of an Illumina HiSeq2000. This enriched endogenous content 24-fold, so that post-capture 30.8–96.5% of reads mapped to the reference genome (Foote et al., 2015).
However, this resulted in a high percentage of duplicate reads (77%) suggesting the libraries contained limited starting template of endogenous DNA (Ávila-Arcos et al., 2015). Thus, the consensus sequence was reduced to 0.02× mean depth of coverage after the removal of duplicate reads. In total, 27,162,031 sites were covered at ≥1×. Analyses of a 41Mb scaffold (KB316842.1) using mapDamage2 (Jónsson et al., 2013) revealed that sequencing reads also exhibited characteristic post-mortem damage patterns (Supporting Information Figure S1). Specifically, an excess of C>T transitions at the 5’ termini as expected from deamination, and the complementary G>A transitions at the 3’ termini. Therefore, all downstream analyses on the type D sequence were restricted to transversions.

Read trimming, mapping, filtering and repeat-masking was conducted as per Foote et al. (2016), with the exception that the previous study masked reads with a combined coverage of 200×, here we masked regions of low (less than a third of the mean) and excessive (more than twice the mean) combined coverage; regions of poor mapping quality (Q<30); and regions called as Ns in the reference sequence, all assessed using the CALLABLELOCI tool in GATK (McKenna et al., 2010; DePristo et al., 2011) and subsequently masked using BEDtools (Quinlan & Hall, 2010). Sites were further filtered to include only autosomal regions, except where stated otherwise, and only sites with a base quality scores >30 were used in all downstream analyses.

Changes in the cluster generation chemistry from bridge amplification to exclusion amplification has been reported to result in increased leakage of reads between indexes in pooled samples on a lane of the new Illumina sequencing platforms, which include the HiSeq4000 (Sinha et al., 2017). To assess the extent of leakage between indexes we mapped the sequencing reads from two HiSeq4000 lanes to a haploid killer whale mitochondrial genome (KF164610.1). Mitochondrial genomes had previously been sequenced using the 454 Life Sciences (Roche) and Illumina HiSeq2000 sequencing platforms (Morin et al., 2010; 2015) and were used as a reference panel. The consensus sequence generated from the HiSeq4000 sequencing reads for this study, were 100% identical for the same individuals in the reference panel (Morin et al., 2010; 2015). We then quantified contamination from leaked reads based on the protocol for assessing the extent of human contamination in Neanderthal sequenced data (Green et al., 2010). We inspected reads from a North Pacific ‘offshore’ killer whale, which mapped to the mitochondrial genome at 695× mean coverage (±72 S.E.) after filtering to remove low-quality bases (Q<30). At sites where the nucleotide was known to be
private to the offshore sequence, we checked the counts of reads that concurred with the 454 Life Sciences (Roche) sequence generated from long-range PCR amplicons for the same individual (Morin et al., 2010), and the counts of reads that matched sequences of one or more of the other individuals pooled on the same HiSeq4000 lane. Counts of the mismatch alleles were uniformly low (mean < 0.5% of reads per site). Our results therefore concur with recent studies (Owens et al., 2018; van der Valk et al., 2018) that the rate of index swapping is low on the new Illumina platforms, and provide confidence that leakage between indexes did not greatly influence the inferred genotype likelihoods or the inferred genetic relationship among individuals.

2.3 | Principal component analysis

The relationship of the samples in the global dataset to the killer whale ecotypes were explored using PCAngsd, a Principal Component Analysis (PCA) for low depth next-generation sequencing data using genotype likelihoods (GLs), thereby accounting for the uncertainty in the called genotypes that is inherently present in low-depth sequencing data (Meisner & Albrechtsen, 2018). We restricted the analyses to autosomal scaffolds ≥10Mb in length, which accounted for 1.5 Gb (~64%) of the genome (prior to masking). First, we estimated covariance of the 46 modern samples, which included five samples each for resident, transient, types B1, B2 and C. After pruning SNPs to reduce linkage, a total of 225,281 SNPs were considered in this analysis. The eigenvectors from the covariance matrix were generated with the R function 'eigen', and significance was determined with a Tracy–Widom test (Tracy & Widom, 1994; Patterson, Price & Reich, 2006) performed in the R-package AssocTest (Wang et al., 2017) to evaluate the statistical significance of each principal component identified by the PCA. To reduce the influence of variable sample sizes among populations, we then repeated the analyses with a subset of 25 samples, removing four samples each from the resident, transient, types B1, B2 and C, and removing the Norwegian sample, which belongs to the same metapopulation as the Icelandic sample (Foote et al., 2011).

To investigate the relationship between the type D morph and the other killer whale populations, a further PCA was constructed using the single read sampling approach implemented in ANGSD (Korneliussen, Albrechtsen & Nielsen, 2014). A random base was sampled from each position at which all samples were covered at ≥1× coverage to remove bias caused by differences in sequencing depth. Transitions were then removed due to the confounding effect of DNA damage patterns caused by deamination appearing as C→T.
transitions, and the corresponding reverse complement A→G in the sequence data. Random bases were therefore sampled from a total of 6,565 transversions.

### 2.4 Individual assignment and admixture analyses

An individual-based assignment test was performed using NGSadmix (Skotte, Korneliussen & Albrechtsen, 2013), a maximum likelihood method that bases its inference on genotype likelihoods. As for the PCA analysis, we ran the NGSadmix twice, once with the 46 modern samples (i.e. excluding the type D museum specimen) and once with the subset of 25 samples. As above, analyses were restricted to autosomal scaffolds ≥10Mb in length. NGSadmix was run with the number of ancestral populations $K$ set from 1–10. For each of these $K$ values, NGSadmix was re-run five times for each value of $K$, and with different seeds to ensure convergence. SNPs were further filtered to include only those covered in at least 25 individuals with a probability of $P < 0.000001$ of being variable as inferred by the likelihood ratio test and removing sites with a minor allele frequency of 0.05, so that singletons were not considered. Finally, SNPs were pruned to account for linkage, resulting in the analyses being based on 290,309 variant sites.

### 2.5 Distance-based phylogenetic inference

The genetic relationships among individuals within the dataset were further reconstructed with ngsDist (Vieira et al., 2015) using distance-based phylogenetic inference based on pairwise genetic distances. ngsDist takes genotype uncertainty into account by avoiding genotype calling and instead uses genotype posterior probabilities estimated by ANGSD. A block-bootstrapping procedure was used to generate 100 distance matrices, obtained by repetitively sampling blocks of 500 SNPs from the original data set of 6,974,134 SNPs. Pairwise genetic distances among the 5× coverage genomes were visualised as a phylogenetic tree using the distance-based phylogeny inference program FastME 2.1.4 (LeFort, Desper & Gascuel, 2015).

### 2.6 Pairwise sequentially Markovian coalescent

We used seqtk (https://github.com/lh3/seqtk) to combine 32 haploid male X-chromosome scaffolds of >1Mb each and totalling 91Mb, to construct pseudo-diploid sequences. The PSMC model estimates the Time to Most Recent Common Ancestor (TMRCA) of segmental blocks of the genome and uses information from the rates of the coalescent events to infer $N_e$ at a given time, thereby providing a direct estimate of the past demographic changes of a population.
(Li & Durbin, 2011). The method has been validated by its successful reconstructions of demographic histories using simulated data and genome sequences from modern human populations (Li & Durbin, 2011). A consensus sequence of each bam file was then generated in fastq format sequentially using the SAMtools mpileup command with the –C50 option to reduce the effect of reads with excessive mismatches (Li et al., 2009); bcftools view --c to call variants; lastly, vcfutils.pl vcf2fq to convert the vcf file of called variants to fastq format. Pairs of fastq files were then merged using seqtk and PSMC inference carried out using the recommended input parameters for human autosomal data (Li & Durbin, 2011), i.e. 25 iterations, with maximum TMRCA ($T_{\text{max}}$) = 15, number of atomic time intervals ($n$) = 64 (following the pattern $1^*4 + 25^*2 + 1^*4 + 1^*6$), and initial theta ratio ($r$) = 5. Plots were scaled to real time as per (Li & Durbin, 2011), assuming a generation time of 25.7 years (Taylor et al., 2007) and a neutral mutation rate of the X-chromosome ($\mu_X$) derived as $\mu_X=\mu_A[2(2+\alpha)]/[3(1+\alpha)]$, assuming a ratio of male-to-female mutation rate of $\alpha = 2$ (Miyata et al., 1987) and an autosomal mutation rate ($\mu_A$) of $2.34 \times 10^{-8}$ substitutions/nucleotide/generation (Dornburg et al., 2011). This gave us an estimated $\mu_X = 2.08 \times 10^{-8}$ substitutions/nucleotide/generation. Only males were used in these analyses, which included transient, resident, Antarctic types B1 and C as our focal ecotypes (our 5× coverage type B2 genome sequence was generated from a female); and from our global dataset we included the sequences of samples from Gabon, Gibraltar, New Zealand, North Pacific offshore ecotype, Eastern Tropical Pacific (ETP)-Clipperton Island, Iceland, Gulf of Mexico, Brazil, Southern Ocean, SW. Australia, Chatham Islands, Crozet Archipelago, Hawaii, ETP–Mexico, and W. Australia.

2.7 | Inferring admixture from D- and F-statistics

To investigate whether ecotype pairs evenly shared derived alleles with outgroup populations, or whether one ecotype shared an excess of derived alleles with outgroups suggesting either recent shared ancestry or introgression, we estimated the $D$-statistic (Green et al., 2010) for various combinations of ecotypes and outgroups. For example, if the sympatric North Pacific resident and transient ecotypes are considered to be the in-group, and $X$ is a global outgroup sample, the test can be used to evaluate if the data are inconsistent with the null hypothesis that the tree (((resident, transient), X), dolphin) is correct and that there has been no gene flow between $X$ and either resident or transient, or any populations related to them. The definition used here is from Durand et al. (2011):

$$D = \frac{n_{ABBA} - n_{BABA}}{(n_{ABBA} + n_{BABA})}$$
Where in the tree given above, \(n_{ABBA}\) is the number of sites where \(\text{resident}\) shares the ancestral allele with the dolphin, and \(\text{transient}\) and \(X\) share a derived allele (ABBA sites); and, \(n_{BABA}\) is the number of sites where \(\text{transient}\) shares the ancestral allele with the dolphin, and \(\text{resident}\) and \(X\) share a derived allele (BABA sites). Under the null hypothesis that the given topology is the true topology, we expect an approximately equal proportion of ABBA and BABA sites and thus \(D = 0\). Hence a test statistic that differs significantly from 0 provides evidence either of gene flow, or that the tree is incorrect. The significance of the deviation from 0 was assessed using a \(Z\)-score based on jackknife estimates of the standard error of the \(D\)-statistics. This \(Z\)-score is based on the assumption that the \(D\)-statistic (under the null hypothesis) is normally distributed with mean 0 and a standard error achieved using the jackknife procedure. The tests were implemented in ANGSD and performed by sampling a single base at each position of the genome to remove bias caused by differences in sequencing depth at any genomic position. An error in our script reversed the sign of the value of \(D\) in a previous study (Foote et al., 2016), thus results differ in the direction of gene flow between that study and this, but do not change the conclusions drawn in that study, \(i.e.\) that some ecotypes are admixed.

The \(f_3\)-statistic is based on the quantification of genetic drift (change of allele frequencies) between pairs of populations in a tree using variance in allele frequencies (Reich et al., 2009; Patterson et al., 2012; Peter, 2016). The \(f_3\)-statistic can provide evidence of admixture, even if gene flow events occurred hundreds of generations ago (Patterson et al. 2012). These tests are of the form \(f_3(A;B,C)\), where a significantly negative value of the \(f_3\) statistic implies that population \(A\) is admixed (Patterson et al., 2012). \(f_3\)-statistics were computed using the estimators described in Patterson et al. (2012), obtaining standard errors using a block jackknife procedure over blocks of 1,000 SNPs.

When a potentially admixed population is identified the admixture proportions can be estimated using the ratio of \(f_4\)-statistics. The expected value of the statistic \(f_4(A,B;C,D)\) would be zero if we see no overlap in the paths of allele frequency changes between \(A\) and \(B\), and between \(C\) and \(D\) through the tree. The expected value of the statistic \(f_4(A,B;C,D)\) will be negative and significantly different from zero if allele frequency changes between \(A\) and \(B\) and between \(C\) and \(D\) take paths in the opposite direction along a shared edge within the tree; or positive and significantly different from zero if the drift between \(A\) and \(B\) and between \(C\) and \(D\) share overlapping paths in the same direction along an edge within the tree. The \(f_4\)-statistic is not sensitive to post-admixture drift and can provide evidence of admixture, even if gene
flow events occurred hundreds of generations ago (Patterson et al., 2012). To better identify the source populations that have admixed with the North Pacific transient ecotype we estimated $f_4$(NZ, pilot whale; $X$, transient) and compared these with an estimate of $f_4$(NZ, pilot whale; resident, transient). Patterson et al. (2012) defined the $f_4$-ratio test as:

$$
\frac{f_4(C, O; X, B)}{f_4(C, O; A, B)}
$$

Where A and C are a sister group, B is sister to (A,C), X is a mixture of A and B, and O is the outgroup. This ratio estimates the ancestry from A, denoted as $\alpha$, and the ancestry from B, as $1-\alpha$. It important to remember that neither the transient nor resident represent unadmixed lineages (see results), however, they do appear from past studies to represent the two most differentiated populations in the North Pacific (Hoelzel et al., 2007; Parsons et al., 2013; Morin et al., 2010, 2015). $F_4$-statistics were computed using the estimators described in Patterson et al. (2012), obtaining standard errors using a block jack-knife procedure over blocks of 1,000 SNPs.

2.8 | Detecting archaic tracts

The prevalence of private alleles in Antarctic types B1, B2 and C, as identified by the PCA, suggested potential archaic introgression from a now extinct (or otherwise unsampled) killer whale lineage or sister taxon. Archaic tracts with a distinctly older TMRCA than the genome-wide average can be identified even without an archaic reference genome. Private alleles resulting from de novo mutation along the branch to the Antarctic should be approximately randomly distributed across the genome, whereas tracts introgressed from a divergent lineage after vicariance of the Antarctic types from other populations, or differentially sorted from structure in an ancestral population will contain clusters of private alleles, the density of which will depend upon the divergence time of the introgressing and receiving lineages (Racimo et al., 2015). Since tract length is a function of recombination rate and time, tracts from ancestral structure are expected to be shorter than recently introgressed tracts due to the action of recombination (Racimo et al., 2015).

We therefore set out to screen for genomic tracts of consecutive or clustered private alleles in the Antarctic types. To ensure the results are comparable despite variation between samples in coverage at some sites, we randomly sampled a single allele at each site from each diploid modern genome. For the outgroup we used all variants found in a dataset consisting of the following widely distributed non-Antarctic samples: Gabon, Gibraltar, New Zealand, North Pacific offshore, resident and transient ecotypes, ETP–Clipperton Island, Iceland, Norway,
Newfoundland, Hawaii, ETP–Mexico, Scotland and W. Australia; *i.e.* samples that show no evidence of admixture with the Antarctic types in either the PCA or NGSadmix analyses. For the ingroup we used *type B2*, which from the $f_3$-statistics appeared to be the least admixed of the Antarctic types. Thus, we consider only variants found in *type B2*, which are not found in the widely distributed 14 non-Antarctic samples listed above.

We then used a Hidden Markov Model (HMM) to classify 1 kb windows into ‘non-archaic’ and ‘archaic’ states based on the density of private alleles (*Skov et al.*, 2018). The background mutation rate was estimated in windows of 100 kb, using the variant density of all variants in non-Antarctic populations. We then weighted each 1 kb window by the proportion of sites not masked by our RepeatMasker and CallableLoci bed files. The HMM was trained using a set of starting parameters based on those used for humans (*Skov et al.*, 2018). We trained the model across five independent runs, varying the starting parameters each time to ensure consistency of the final parameter input. Posterior decoding then determines whether consecutive 1 kb windows change or retain state (‘archaic’ or ‘non-archaic’) dependent upon the posterior probability.

### 3 | RESULTS

Our results highlight that the distinctiveness of the killer whale ecotypes as described in the Introduction section, reflects their demographic and evolutionary histories, which include deep ancestral splits masked by more recent admixture. The latter confounding the inference of the relationships among these populations as a simple bifurcating tree-like model.

#### 3.1 | Genome sequences

Short-read sequence data were generated for 27 individuals, resulting in a mean coverage of 5× coverage of the autosomal region of the killer whale genome for the 26 modern samples, and 27Mb covered at ≥1× from a 62-year old museum specimen of the *type D* subantarctic morph. For some analyses, these data were combined with 20 previously sequenced 2× coverage genomes from the North Pacific *transient* and *resident* ecotypes, and Antarctic types *B1, B2* and *C* (*Foote et al.*, 2016) and ~20× coverage RAD-seq data (*Moura et al.*, 2015).

#### 3.2 | Principal Component Analysis: genetic variation segregates in Antarctic types and resident ecotype
In a PCA that included five samples per ecotype, the Antarctic types (B1, B2 and C) separated out from all other killer whales along PC1 (Figure 1b), which explained 24.2% of the variation (Supporting Information Figure S2a). The resident ecotype formed a distinct cluster which separated out from a third cluster containing the transient ecotype and all other samples along PC2 (Figure 1b), explaining 9.7% of the variation (Supporting Information Figure S2a). Both first and second components were statistically significant: $P < 0.001$. This result was replicated when published RAD-seq data for a sub-Antarctic Marion Island sample were included (Supporting Information Figure S3a). The transient ecotype partially segregates from other samples along PC4, which explains 2.6% of the total genetic variation (Supporting Information Figure S4).

Uneven sampling of different demes can influence the inference of population clusters in admixture and PCA plots (McVean, 2009; Gilbert, 2016; Lawson, van Dorp & Falush, 2018). For example, when the closely related Norwegian and Icelandic samples are both included in the PCA, they segregate from the other samples along PC2 (Supporting Information Figure S5). After reducing the dataset used in the PCA to one sample per population to reduce this bias, differences between the Antarctic types and all other killer whales continue to explain the greatest and only significant ($P < 0.001$) component of variation in the data (Figure 1c; Supporting Information Figure S2b). This pattern remains even when just a single type B1 sample (i.e. no B2 or C samples) is included, and likewise for single B2 and C samples, albeit with less variation (~11%) explained by PC1 (Supporting Information Figure S6). Reducing the dataset to one sample per population results in a change in clustering along PC2, along which samples are distributed, to some extent, reflecting a cline from the North Atlantic to the North Pacific, but with the resident and offshore samples intermediate between Atlantic and Pacific samples (Figure 1c). A further PCA identified the type D morphotype (Pitman et al., 2010) as being genetically intermediate between Southern Ocean and North Atlantic populations (Supporting Information Figure S7).

3.3 | Individual assignment and admixture analyses support PCA inference

The results of the PCAs are reflected in the admixture plots inferred by NGSadmix (Figure 1d,e). The uppermost hierarchical level of structure, inferred from the greatest step-wise increase in log likelihood ($\Delta K$; Evanno, Regnaut, & Goudet, 2005) identified two clusters (Supporting Information Figure S8). As in the PCA, one cluster consisted of Antarctic types B1, B2 and C, the other a mostly homogenous cluster of all other killer whales, albeit with ‘Antarctic’ ancestry detected in some southern hemisphere samples (Figure 1d,e). When
multiple samples per ecotype are included, we find the second highest $\Delta K$ from $K=2$ to $K=3$ clusters, in which the North Pacific resident ecotype forms a discrete cluster (Figure 1d).

PCA and STRUCTURE-like admixture models use similar information and generate similar axes of variation (Patterson, Price & Reich, 2006; Lawson, van Dorp & Falush, 2018). Both methods are likely to identify the samples with the greatest population-specific drift that therefore share rare derived alleles or have lost ancestral alleles from standing variation, as the major axes of structure (Lawson et al., 2018). Accordingly, both PCangsd and NGSadmix identified the uppermost hierarchical level of structure within our dataset as being between the Antarctic types and all other killer whales (Figure 1). The spatial distribution of samples within the PCA plot can be considered as being representative of the mean pairwise coalescent times between each pair of samples (McVean, 2009). Changes in frequency or the loss of neutral alleles through population specific drift will result in more recent mean coalescence among individuals, thereby causing them to cluster together and segregate from other populations along the axes of the PCA (McVean, 2009). Our results are therefore consistent with previous findings of a shared demographic history of the Antarctic types that included a shared population bottleneck and substantial drift (Morin et al., 2015; Foote et al., 2016). However, our finding that this pattern in the PCA is retained when just a single Antarctic sample is included (Figure S7), indicates this pattern is not just driven by the shared loss of standing variation in the Antarctic types, but that alleles explaining a significant proportion of the observed genetic variation coalesce within that single sample in those analyses. This suggests that there is are a large number of alleles private to the Antarctic types contributing towards the pattern of global genetic variation in killer whales. It should be noted that at higher values of $K$ further structuring is identified within our dataset; for example, at $K=7$ Antarctic types $B1$, $B2$ and $C$ form three discrete clusters as per Foote et al. (2016). However, due to our sampling scheme, our focus in this study was not to identify regional structuring, but to identify the major axes of global structure and to infer the underlying processes.

3.4 ‘Archaic’ tracts in Antarctic types suggest ancient admixture with an ancient ‘ghost’ population

The HMM method for detecting archaic tracts based on private allele density (Skov et al., 2018) inferred 1,897 tracts totalling 8,119 kb of the tested 41 Mb scaffold as archaic in Antarctic type $B2$ with a posterior probability of $\geq 0.5$ (Supporting Information Figure S9). Thus, up to 21.6 % of the genomic region analysed was inferred to be potentially comprised of
archaic ancestry. However, a proportion of these windows inferred to be in the archaic state with a posterior probability of $\geq 0.5$ may be false positives. Considering windows inferred as archaic with posterior probabilities of $\geq 0.8$ identified 18 archaic tracts totalling 224 kb.

The emission probabilities of the HMM are modelled as Poisson distributions with means of $\lambda_{\text{Archaic}} = \mu \cdot L \cdot T_{\text{Archaic}}$ for the archaic state and $\lambda_{\text{Ingroup}} = \mu \cdot L \cdot T_{\text{Ingroup}}$ for the non-archaic (or ingroup) state (Skov et al., 2018), where $L$ is the window size (1000 bp) and $\mu$ is the mutation rate ($2.38 \times 10^{-8}$; Dornburg et al., 2011). This allows us to estimate the mean TMRCA of the archaic and ingroup windows with the corresponding segments in the outgroup dataset. The TMRCA between the archaic tracts within type B2 and the corresponding genomic regions in the outgroup is a Poisson distribution around a mean of 9,786 generations (~251 KYA, assuming a generation time of 25.7 years; Taylor et al., 2007). The estimated TMRCA between the non-archaic tracts within type B2 and the corresponding genomic regions in the outgroup is a Poisson distribution around a mean of 2,429 generations (~62 KYA). Thus, the genome of type B2 appears to be admixed, comprising of approximately 80% ancestry that coalesces with the ancestry of the outgroup during the previous glacial period (Marine Isotope Stage 5), and approximately 20% ancestry derived from an unsampled lineage that coalesces with the ancestry of the outgroup during an earlier glacial cycle (Marine Isotope Stage 8).

Tracts inferred to be in the archaic state with a posterior probability of $\geq 0.5$ were on average between 5 and 6 windows long, i.e., between 5 and 6 kb; an order of magnitude shorter than introgressed archaic tracts in non-African humans (Skov et al., 2018). Considering just the tracts called as archaic with posterior probabilities of $\geq 0.8$, the average tract length was 12-13 kb. The estimation of the time of introgression from tract length is dependent upon recombination rate ($r$), which has not yet been estimated for killer whales. However, assuming a constant value of $r$ approximated to the mean $r$ estimated for the human genome of $1.2 \times 10^{-8}$/bp (1.2 cM/Mb; Dumont & Payseur, 2008) places the age of tracts 5.5 kb long to approximately 14,000 generations ago, i.e. older than $T_{\text{Archaic}}$ (Supporting Information Figure S10). A recombination rate of $5.0 \times 10^{-8}$/bp (5.0 cM/Mb) would be required for such short tract lengths to have introgressed during the last 2,500 generations, i.e. close to $T_{\text{Ingroup}}$ (Supporting Information Figure S10). Considering tracts of 12.5 kb length suggests a time of introgression between $T_{\text{Archaic}}$ and $T_{\text{Ingroup}}$ of approximately 7,000 generations (assuming $r = 1.2 \times 10^{-8}$/bp). Assuming a recombination rate for killer whales in the range of humans thus suggests a scenario different from the recent introgression from Neandertals into the lineage of
non-African humans. Instead, the source of archaic ancestry tracts in type B2 killer whales is better explained by ancestral population structure. This therefore requires a scenario in which these tracts were the minor component of the ancestry (i.e., the lineage that contributed less to the gene pool, see Schumer et al., 2018) of an admixed ancestral killer whale population between $T_{\text{Archaic}}$ and $T_{\text{Ingroup}}$, and this ancestry was therefore already being broken up by recombination prior to $T_{\text{Ingroup}}$ (Figure 2).

A PCA plot of SNPs occurring within the 1 kb windows inferred by the HMM as being in the archaic state (at posterior probability >0.5) highlighted the role of these archaic tracts in contributing to the major axis of structure in our global dataset (Supporting Information Figure S11). It also indicates variation among types B1, B2 and C in the sharing of variants within these tracts. A PCA estimated from the non-archaic tracts (not shown) generated similar PCs and eigen values to Figure 1c and so the differentiation of the Antarctic types is not driven solely by the archaic tracts.

3.5 | PSMC suggests an early split of transients, and ancestral vicariance and admixture in Antarctic types

To better understand the chronology of the divergence of killer whale ecotypes, we employed a method that drew inference from the distribution of the lengths of shared Identity-by-State (IBS) tracts to investigate coalescence rates through time. We created pseudo-diploid sequences by combining the phased haploid non-pseudoautosomal X-chromosome sequences from two different males and used PSMC (Li & Durbin, 2011) to estimate changes through time in the coalescent rate between the two X-chromosome haplotypes. The $y$-axis of a PSMC plot is driven by both changes in population structuring and demography and is more accurately interpreted as an estimate of the inverse of the rate of coalescence at any point in time represented along the $x$-axis (Mazet et al., 2016). Pseudo-diploids comprised of two haploid male X-chromosome sequences can therefore be used to infer the approximate population split time between two populations (Li & Durbin, 2011). When populations diverge and all gene flow between them ceases, the accumulation of new mutations and loss of diversity through drift will be population specific (Figure 3a). Population divergence therefore manifests itself in the pseudo-diploid sequence as heterozygote sites that break up long homozygous tracts from which more recent coalescent events are inferred. This results in an upsweep along the $y$-axis of the PSMC plot approximately at the point of cessation of gene flow (Li & Durbin, 2011). Post-divergence migration between the two demes being compared can result in a more
recent coalescence of post-divergence mutations, shifting the upsweep closer to the present along the x-axis (Cahill et al., 2016).

Applying this approach, we infer decreasing coalescence from the upsweep in estimated $N_e$ from 200-300 KYA between the transient and all tested populations (Figure 3b). In contrast, coalescence does not appear to decrease between the resident and these same populations until approximately 100 KYA (Figure 3b). In other words, the resident shares a higher proportion of longer IBS tracts within the X-chromosome with the tested global samples, representing more recent recombination events, than the transient does with those same populations. Or to put it another way, the mean TMRCA of the X-chromosome is older between the transient ecotype and the populations tested here, than between the resident ecotype and those same populations. We interpret this as an earlier matrilineal fission and divergence from these globally distributed samples by the ancestor of the present-day transient ecotype, and a later founding of the ancestor of the present-day resident ecotype. This is consistent with earlier estimates of TMRCA based on mitochondrial genomes (Morin et al., 2015) and the inferred timing of founder bottlenecks based on nuclear genomes (Foote et al., 2016).

Comparing coalescent patterns of the X-chromosome of a type B1 male and a type C male with the global dataset we find that the B1-global pseudo-diploid plots follow a similar trajectory to the transient-global plot, whereas the C-global pseudo-diploid plot upsweep suggests a more recent decrease in coalescence with the global samples (Figure 3c). Thus, despite the covariance of allele frequencies and resulting clustering in the PCA and admixture plots (Figure 1), type B1 and type C differ in their sharing of shorter IBS segments with a TMRCA >100 KYA. An upsweep of inferred $N_e$ in plot of the types B1 and C pseudo-diploid commencing between 200-300 KYA, stalls at approximately 90 KYA and declines between 90-50 KYA, before increasing again (Figure 3c). The increase in coalescence, estimated between 50-90 KYA, implies a period of admixture between types B1 and C. The bootstrap plots illustrate the variation in ancestry across the X-chromosome (Figure 3d).

3.6 | Comparing mitochondrial DNA (mtDNA) and nuclear DNA (nuDNA) tree topologies
A distance-based tree estimated from pairwise genetic differences is only partially concordant with the mitochondrial DNA tree (Supporting Information Figure S12). Thus, in some cases genetic variation of the nuclear genome is shared among samples reflecting the matrilineal fission process that drives divergence in social and population structure in most killer whale populations studied to date (Ford, 2009); in other cases, geographically proximate samples
with divergent mtDNA haplotypes cluster in the nuclear tree, suggesting a role for long-range matrilineal dispersal and subsequent gene flow in shaping patterns of nuclear genome diversity.

3.7 | D-statistics indicate Antarctic types B1, B2 and C differ in their sharing of derived alleles with outgroup populations

The D-statistic (Green et al., 2010) considers a tree-like relationship among four populations, e.g. (ecotype1, ecotype2; X, dolphin), and estimates whether X shares an excess of derived alleles with one of the two ecotypes in the ingroup. Significant sharing of derived alleles between an in-group and X indicates either introgression from X (or a closely related population) into one ecotype, but not the other; or that the tree topology is incorrect, and X belongs in the in-group. Estimation of D(type B2, type C; X, dolphin) found that 19 out of 23 tests were considered significant based on Z-score > 3, and that these 19 samples shared an excess of derived alleles with type C relative to type B2 (Figure 4a). A similar result was obtained when type B2 was swapped for type B1, i.e. D(type B1, type C; X, dolphin); in this test type C shares a significant excess of derived alleles with all outgroups (X), except the sample from the Crozet Archipelago. There was no significant difference between type B1 and type B2 in the sharing of alleles derived in any of the outgroup samples (X).

3.8 | D-statistics indicate Pacific resident and transient ecotypes differ in their sharing of derived alleles with outgroup populations

Estimation of D(transient, resident; X, dolphin) found that while none of the global outgroup samples shared a significant excess of derived alleles with the transient ecotype, a widely geographically distributed set of 16 out of 24 samples shared a significant (Z-score < -3) excess of derived alleles with the resident ecotype (Figure 4b). This may indicate that the topology (transient, resident; X, dolphin) is incorrect, and that in these 16 tests (resident, X) is the correct in-group, which would be consistent with the hypothesis of secondary genetic contact between the transient and resident ecotypes (Foote et al., 2011). We therefore compared D(X, resident; transient, dolphin) and D(X, transient; resident, dolphin) to assess this possibility (Figure 4c).

When we consider alleles purportedly derived in the transient (i.e. where X and resident are the ingroup) there is no significant sharing of excess derived alleles between any population represented by (X) and the transient ecotype. However, when X was a non-North Pacific sample, there was significant sharing of derived alleles between the resident and transient ecotypes (Figure 4c). When we consider X and the transient as the ingroup (i.e. alleles...
purportedly derived in the resident) the North Pacific offshore, ETP–Clipperton Island, New Zealand, Gibraltar, Gabon and Scotland samples all shared a significant excess of derived alleles with the resident ecotype (Figure 4c). These same populations generated the most strongly negative D-statistics in the test $D(\text{transient, resident; } X, \text{ dolphin})$ (Figure 4b) and share a more recent common maternal ancestor with the resident than the transient ecotype based on mitochondrial genome phylogeny (Morin et al., 2015; Supporting Information Figure S12). We therefore interpret these results as a further indication that the resident ecotype diverged more recently from these six populations than it did from the transient, but that gene-flow between the resident and transient has subsequently occurred, most likely within the North Pacific. A comparison of $D(\text{transient, resident; } X, \text{ dolphin})$ and $D(\text{transient, offshore; } X, \text{ dolphin})$ indicates correlated (Pearson’s correlation coefficient: $r^2 = 0.9609$, $p < 0.00001$) sharing of derived alleles between the resident and $X$, and between the offshore and $X$ (Supporting Information Figure S13) supports this inference of recent shared ancestry between the resident and offshore ecotypes.

3.9 | F-statistics indicate admixture between transient and resident lineages

The $f_3$-statistic is based on the quantification of covariance of allele frequencies (often referred to as shared drift) between pairs of populations in a tree using variance in allele frequencies (Reich et al., 2009; Patterson et al., 2012). To identify admixed ecotypes, we estimated $f_3$-statistics of the form $f_3(\text{ecotype; } X, Y)$. Significantly negative $f_3$-statistics indicate varying levels of admixture between the ecotype and $X$, and between the ecotype and $Y$, so the estimate of allele frequency differences between the ecotype and $X$ are negatively correlated with the differences in allele frequencies between the ecotype and $Y$. $F_3$-statistics were minimized and significantly ($Z$-score $<-3$) negative when estimating $f_3(\text{transient; } X, Y)$, with the exception of when both $X$ and $Y$ were Antarctic types (Figure 5, Supporting Information Figure S14). This result indicates the transient ecotype is admixed with one or more of the donor populations, or with closely related populations with partly shared derived ancestry with the donor population (see for example the outgroup case in Patterson et al., 2012). The most negative $f_3$-statistics were estimated for tests when Hawaiian and/or Mexican ETP samples represented $X$ and $Y$ (see columns 1 & 2 of the lower diagonal of Figure 5), consistent with the results from PCA (Figure 1c) and D-statistics (Figure 4b,c). Based upon $f_3$-ratio tests, the Hawaiian and Mexican ETP samples shared a higher proportion of transient than resident ancestry (Supporting Information Figure S15).

3.9 | F-statistics indicate drift is greater than any admixture in Antarctic types
$F_3$-statistics were positively maximized when Antarctic type $B2$ was the target ecotype for all tested combinations of $X$ and $Y$ (Figure 5, Supporting Information Figure S14). Positive and non-significant $f_3$-statistics can arise despite admixture, for example, if population specific post-admixture drift in the target population is so large, it masks gene flow from the tested donor populations (Patterson et al., 2012). The extent of drift in type $B2$ is such that it acts as an outgroup in $f_3$-statistics, the same way that African genomes (e.g. Yoruban or Mbuti) are often used as an outgroup when comparing shared ancestry of Eurasian populations as $X$ and $Y$ in studies of human populations (e.g. Seguin-Orlando et al., 2014; Pagani et al., 2016). Therefore, we estimate the highest positive $f_3$-statistics for tests when $X$ and $Y$ are known to originate from closely related populations, for example, $f_3(B2; \text{Norway, Iceland})$ (Figure 5).

4 | DISCUSSION

The present-day major axes of global genetic structure in killer whales are associated with the strongest drift having occurred in populations at the high latitude extremes of the species range. This pattern likely reflects some of the major demographic events in the last tens of millennia of the history of this species. Consistent with the expectations of range expansion theory (Excoffier, Foll & Petit, 2009) we find that populations that have expanded into areas inaccessible during the LGM (e.g. Antarctica and the Northern range limits of the Atlantic and Pacific) or have undergone some other long-range dispersal (e.g. the resident ecotype) have the greatest differentiation from neighbouring populations, indicating they have undergone the greatest drift in allele frequencies. The same high latitude populations show expansion from a small ancestral founder lineage based on the TMRCA of the mitochondrial genome (Morin et al., 2015) and coalescence patterns in the nuclear genome (Foote et al., 2016). Our results therefore expand the model of the evolution of population structure in the North Pacific proposed by Hoelzel et al. (2007), i.e. strong founder effects from ancestral colonising matriline, to partially explain the major axes of structure in killer whales at high latitudes including Antarctica and the Northeast Atlantic (Norway and Iceland). However, despite these commonalities in the demographic and evolutionary histories of high latitude killer whale populations, the Antarctic types $B1$, $B2$ and $C$ stand out as explaining by far the most significant component of global genomic variation in this species.

We find an additional source of genetic variation in the Antarctic types in the form of private alleles clustered within short archaic tracts. The majority of the genome of the Antarctic types coalesces in the shared ancestral bottlenecked population and has a mean TMRCA of approximately 60 KYA with a widely distributed dataset of outgroup samples (Figure 2).
However, we also find short ancestry tracts which have an estimated mean TMRCA of over 200 KYA with corresponding genomic regions in these same outgroup samples (Figure 2). Thus, the genomes of Antarctic killer whales contain ancestry reflecting deeper coalescence during a previous glacial cycle. This is supported by our analysis of changes in coalescence rate through time between X-chromosome haplotypes of Antarctic types B1 and C using PSMC, which although coarse, also indicate two peaks in coalescence, one at ~60 KYA and another >200 KYA. There is further support for an older and younger coalescence among Antarctic types from the TMRCA of mitochondrial genomes (see Supporting Information Figure S16; Morin et al., 2015). We hypothesize that this pattern in the mtDNA phylogeny is the result of replacement of mtDNA haplotype diversity of the Antarctic types during this episode of admixture between 50-90 KYA inferred from the PSMC plot (Figure 2d). The fixation of an introgressed mtDNA haplotype would be dependent upon admixture rate and effective population size (see Posth et al., 2017); high admixture rates and low $N_e$ would be needed to drive the near fixation of the ancestral mtDNA lineage in types B1, B2 and C and the pattern observed in the B1-C pseudo-diploid PSMC plot. Considering all these lines of investigation together, we interpret the results as being indicative of cyclical range expansions and contractions concurrent with the glacial cycles. Antarctic populations would be able to expand their range southwards during inter-glacial periods, increasing genetic differentiation from lower latitude populations, but then would retreat northwards during glacial periods, increasing contact and gene flow with lower latitude populations.

Our finding that the strongest structuring in a global dataset of killer whales is between the ecotypes found around Antarctica (types B1, B2 and C) and all other killer whales counters claims by de Bruyn et al. (2013), that the Southern Ocean provides ‘complete and uninterrupted connectivity’ between Antarctic and Southern Hemisphere killer whales. Despite the apparent homogeneity of the Southern Ocean it harbours geographically structured populations of many species and is a hotbed for adaptation (see examples given in Rogers, 2007 and Moon, Chown & Fraser, 2017). For example, pelagic versus coastal niche, and oceanographic fronts, shape the range, dispersal potential and consequently genetic structuring among Southern Ocean penguin populations (Clucas et al., 2018). Our findings of structure between Antarctic and all other killer whales, in addition to previous findings of structure between Antarctic types B1, B2 and C (Foote et al., 2016) are therefore consistent with patterns in other Antarctic taxa. Our findings make clear that, despite some connectivity, sub-Antarctic and Antarctic killer whale populations should not be conflated.
Ancient vicariance during a previous glacial cycle followed by more recent admixture is also inferred from the ancestry of the sympatric North Pacific mammal-eating *transient* ecotype and the fish-eating *resident* ecotype. The results from the PCA, PSMC and the $D$-statistics indicate more recent mean genome-wide coalescence and greater sharing of derived alleles and longer IBS tracts between the *resident* ecotype and most North Atlantic samples than between the *transient* ecotype and those same populations. We therefore infer that the *resident* ecotype shares a more recent common ancestor with these North Atlantic samples than does the *transient* ecotype. Our finding that alleles derived in the *transient* are shared more commonly with the *resident* than with all non-Pacific samples suggests gene-flow within the North Pacific between *residents* and *transients*, as first inferred from Isolation with Migration analyses (IMa) of microsatellite genotypes by Hoelzel et al. (2007). Whilst this may appear to contradict long-term observations of social isolation between the two ecotypes (Morton 1990; Baird & Dill, 1995; Ford, 2009), gene flow via intermediary populations is supported by the $f_2$-ratio test which identified Eastern Tropical Pacific samples and the *offshore* ecotype as having a mixture of *transient* and *resident* ancestry. Admixture may also be largely ancestral, rather than contemporary. The $f_3$-statistic tests indicate greater drift relative to admixture in the *resident*, compared with the transient. In the PCA (Figure 1b), the segregation of samples along PC2 is driven by coalescence of shared genetic variation within the resident ecotype, *i.e.* lineage-specific drift in the *resident*. Our *resident* samples originate from across the ecotype’s North Pacific range, from Washington State, USA to the Sea of Okhotsk off Eastern Russia. Therefore, the variation segregating in the *resident* ecotype and driving PC2 in Figure 1b must pre-date the separation into the several *resident* sub-populations which have subsequently colonised much of the Pacific rim (Filatova et al., 2018). If this drift in allele frequencies shared among *residents* occurred post-admixture with the *transient* ecotype, it would increase genetic differentiation between the two currently sympatric North Pacific ecotypes. Identifying introgressed haplotype lengths will be an important next step in unravelling this detail of the evolutionary history of killer whale ecotypes.

This pattern of recurrent vicariance and subsequent admixture, likely corresponding to the cyclical expansion and contraction of high latitude habitat during the glacial cycles, contributes towards the genomic heterogeneity within an individual genome. Our results suggest that tracts originating from past vicariance during previous glacial cycles can be numerous and comprise a significant proportion of the genome. It is therefore important to consider such tracts in future analyses. For example, admixture between archaic hominin and modern Eurasian humans can inflate divergence time estimates between African and non-African populations (Alves et al.,
Similarly, the genomes of the Antarctic killer whales represent at least two different demographic histories: the major ancestry component reflects a history in which the Antarctic types appear to be recently derived from other Southern Ocean populations; the minor ancestry reflects an ancient divergence that predates the TMRCA of most other killer whale lineages. Thus, the previous estimated TMRCA of 126–227 KYA (Foote et al., 2016) will be an average of the variation in TRMCA across the genome, thereby ignoring the true complexity of vicariance and admixture among killer whale populations. Furthermore, depending upon the demographic and evolutionary history of these tracts, e.g. if they evolved in a locally adapted ancestral population or if the ancestral effective population size was small, they could harbour adaptive variation associated with this extreme Antarctic climate (as per Racimo et al., 2015) and/or weakly deleterious mutations (as per Harris & Nielsen, 2016; Juric, Aeschbacher & Coop, 2016). Thus, further research into the demographic history of these archaic tracts is warranted.

In summary, the global dataset of genomes analysed here contributes further to the emerging consensus (e.g. Gopalakrishnan et al., 2018; Malinsky et al., 2018; Sinding et al., 2018; Tusso et al., 2018) that the relationship among natural populations is rarely well represented as a bifurcating tree. The evolutionary history of natural populations can include episodic long-range dispersal, population replacement and admixture which greatly transform the distribution of global genetic variation. Furthermore, we highlight the importance of a phenomenon hitherto rarely considered in studies of non-human study organisms, that of archaic tracts within a genomic background with much younger TMRCA. Whilst past studies have simulated gene flow from unsampled ‘ghost’ populations (e.g. Wilson & Bernatchez, 1998; Beerli, 2004; Slatkin, 2005), by using tools previously primarily harnessed for the study of human population dynamics we highlight how genomic data can be leveraged to both identify the genomic regions derived from archaic and ghost populations and quantify their effect on contemporary population structure.
Acknowledgements

The sequencing service was provided by the Genomics Core Facility (GCF), Norwegian University of Science and Technology (NTNU). GCF is funded by the Faculty of Medicine and Health Sciences at NTNU and Central Norway Regional Health Authority. We are grateful to the Museum of New Zealand, Te Papa Tongarewa, Wellington, for providing material from the type D specimen (#1077). A.D.F. was supported by a short visit grant from the European Science Foundation–Research Networking Programme ConGenOmics and by a Swiss National Science Foundation grant (31003A-143393) to L. Excoffier, and by the Welsh Government and Higher Education Funding Council for Wales through the Sêr Cymru National Research Network for Low Carbon, Energy and Environment, and from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 663830. We thank Alison Devault at Arbor Biosciences for support with the WGE experiment and Laurits Skov for advice on the HMM method.
REFERENCES


ecotypes: Insights on the behavior of the rarely observed type D killer whales. Marine mammal science, 32(3), 983-1003.


Figure 1 (a) Sampling locations of the individuals for which twenty-six 5x coverage genomes and one 27Mb partial genome were generated (global dataset). Marker colours are as per the PCA legend. An additional twenty low coverage genomes (ecotype dataset) were used in some analyses, see Foote et al. (2016) for sampling locations. (b) PCA plots of the combined global and ecotype datasets (excluding type D), and (c) when only one 5x coverage genome per population is included. (d) Individual admixture proportions, conditional on the number of genetic clusters (K=2 and K=3), for the combined global and ecotype datasets, and for (K=2) (e) when only one 5x coverage genome per population is included.
Figure 2  Model of hypothesised demographic scenario. Approximately 80% of the ancestry of type B2 has a mean time to most recent common ancestor (TMRCA) of 2,500 generations with the outgroup samples ($T_{\text{Ingroup}}$), whilst 20% of the ancestry of type B2 is in short segments (5-6 kb) with a TMRCA of 10,000 generations with the outgroup samples ($T_{\text{Archaic}}$). We therefore propose lineage sorting of ancestral structure subsequent to $T_{\text{Archaic}}$, so that ancestry represented by red shading was not found in the most recent common ancestor of the outgroup and type B2. After $T_{\text{Ingroup}}$, the red shaded ancestry introgressed into the ancestor of type B2, but not the outgroup, from an unsampled ‘archaic’ source lineage. Figure is adapted from Figure 1a of Skov et al. (2018) and Figure 1 of Racimo et al. (2015). Shading represents the decreasing length of the archaic (red) ancestry tracts through the action of recombination.
Figure 3 Pairwise sequentially Markovian coalescent (PSMC) plots of changes in coalescence rates between haploid male X-chromosomes combined to construct pseudo-diploid X-chromosomes. (a) A schematic diagram of the accumulations of mutations (indicated by circles) in two populations which are initially connected by gene flow, but diverge without further gene flow at the time indicated by the dashed line. A pseudo-diploid comprised of haploid chromosomes from Pop1 and Pop2 can be homozygous for the derived alleles at mutations when there is gene flow and recombination (indicated by yellow stars) between Pop1 and Pop2. Mutations accumulated after the cessation of gene flow will remain private to Pop1 or Pop2 and therefore inferred as heterozygotes in the pseudo-diploid. The accumulation and distribution of heterozygotes in the pseudo-diploid breaks up homozygous tracts which PSMC infers as a decrease in the coalescence rate. Therefore, the exponential upsweep towards infinity on the y-axis of the PSMC plot of a pseudo-diploid genome provides a coarse measure of divergence time. (b) The haploid X-chromosome of a male resident (red) and a male transient (blue) are combined with haploid X-chromosomes of males from the global dataset: Gabon, Gibraltar, New Zealand, North Pacific offshore ecotype, Eastern Tropical Pacific (ETP)–Clipperton Island, Iceland, Gulf of Mexico, Brazil, Southern Ocean, SW Australia, Chatham Islands, Crozet Archipelago, Hawaii, ETP–Mexico, and W. Australia. Each pseudo-diploid is represented by a separate plot. (c) PSMC plots of pseudo-diploid X-chromosome constructed from the haploid X-chromosome of either a male type B1 (aubergine) or a male type C (orange) together with the haploid X-chromosome of a global sample (as for panel b). The plot of the combined B1-C pseudo-diploid X-chromosome is shown in black and shown separately in (d) with the thick black line representing the median and thin grey lines corresponding to 100 rounds of bootstrapping. Inverse coalescence rate is scaled by $2\mu$. 

35
Figure 4  
(a) Samples on the map are coloured by the value of $D(\text{type B2, type C}; X, \text{dolphin})$. The nineteen highest positive statistics were considered statistically significant following correction for multiple testing, based on Z-scores $>3$. This indicates that in these nineteen tests $X$ shares an excess of derived alleles with Antarctic type C relative to type B2. Standard errors are shown as horizontal bars on the markers in the plot to the right.  
(b) Estimates of $D(\text{resident, transient}; X, \text{dolphin})$, in which the sixteen most negative statistics were considered statistically significant following correction for multiple testing, based on Z-scores $<-3$. This indicates that in these sixteen tests $X$ shares an excess of derived alleles with the North Pacific resident ecotype relative to the North Pacific transient ecotype. Standard errors are shown as horizontal bars on the markers in the plot to the right.  
(c) A comparison of $D(X, \text{resident}; \text{transient, dolphin})$ and $D(X, \text{transient}; \text{resident, dolphin})$. Negative values along the x-axis indicate samples which shared an excess of derived alleles with the transient ecotype. Negative values along the y-axis indicate samples which shared an excess of derived alleles with the resident ecotype.
Figure 5  $F_3$-statistics of the form $f_3(\text{ecotype}; X, Y)$, showing $f_3(\text{type B2}; X, Y)$ in the upper diagonal and $f_3(\text{transient}; X, Y)$ in the lower diagonal. The negative values for $f_3(\text{transient}; X, Y)$ indicate that the transient ecotype is highly admixed by X and/or Y, or population(s) closely related to them. The positive values for $f_3(\text{type B2}; X, Y)$ indicate that post-admixture drift from X and Y is greater than any admixture with X or Y, with the exception of type B1 or type C.
Se as coisas nos reduzem simplesmente ao nada, do nada simplesmente temos que partir.

*Trem Bala* | Antônio Cícero, Waly Salomão and João Bosco