

Comparative genomic approaches to human evolutionary history

Von der Fakultät für Lebenswissenschaften
der Universität Leipzig

genehmigte

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
Dr. rer. nat

Vorgelegt von

Master of Arts
Alexander Cagan

geboren am 31.10.1987 in London (UK)

Dekan: Prof. Dr. Tilo Pompe

Gutachter: Prof. Dr. Svante Pääbo

Prof. Dr. Leif Andersson

Tag der Verteidigung: 22.09.2017

Bibliographical Summary

Cagan, Alexander TJ

Comparative genomic approaches to human evolutionary history

Fakultät für Biowissenschaften, Pharmazie und Psychologie

Universität Leipzig

Dissertation

177 Seiten, 342 Literaturangaben, 38 Abbildungen, 150 Tabellen

Understanding the success of the human species is central to evolutionary anthropology. While we share many traits with our relatives the great apes, only humans migrated to all corners of the earth and domesticated other species, leading to the emergence of complex societies. Investigations into human genomes have shown that they are a rich source of information for insights into our past. However, for a complete understanding of human evolution it is necessary to look beyond our own genomes. This thesis is about using comparative genomics to place human evolution within a wider context by studying adaptation in our closest living relatives and in the species that we domesticated.

In the first study, I investigate the genetic changes involved in the earliest stages of dog domestication. Using a global sample of dog and wolf genomes I identify regions that are highly diverged between these species. I find that selection in the initial stages of dog domestication likely involved genes involved in the fight-or-flight response, advancing our understanding of this process.

In the second study, I look for commonalities in the genetic changes that occurred during animal domestication across species. I compare genome sequences from experimentally and historically domesticated species. I identify genes and variants that may underlie the phenotypic changes that occurred during domestication. I find evidence of biological pathways that appear to always be involved in the domestication process.

In the third study, I characterise the signatures of natural selection in all major *Hominidae* lineages using population genomic data. I find that most signatures of positive selection are species specific, although some loci appear to be selected across several lineages. I determine that the efficacy of selection varies between species and is significantly correlated with long-term effective population size. These results contribute to a more complete understanding of human evolution.

This thesis is based on the following manuscripts:

1. **Cagan A** & Blass T. (2016) Identification of genomic variants putatively targeted by selection during dog domestication. *BMC Evolutionary Biology*,16:1.
2. **Cagan A**, Albert FW, Plyusnina I, Trut L, Renaud G, Romagné F, Wiebe V, Kozhemjakina R, Gulevich R, Trapezov O, Yudin N, Alekhina T, Aitnazarov R, Trapezova L, Herbeck Y, Schöneberg T, Pääbo S. Genes and pathways selected during animal domestication. Submitted to *eLife*.
3. **Cagan A**, Theunert C, Laayouni H, Santpere G, Pybus M, Casals F, Prüfer K, Navarro A, Marques-Bonet T, Bertranpetit J, Andrés AM. (2016). Natural Selection in the Great Apes. *Molecular Biology and Evolution*, 33:3268-3283.

CONTENTS

Summary	1
Zusammenfassung	8
Chapter 1	16
Identification of genomic variants putatively targeted by selection during dog domestication	
Chapter 2	38
Genes and pathways selected during animal domestication	
Chapter 3	73
Natural Selection in the Great Apes	
References	149
Acknowledgments	152
Curriculum Vitae	153
Declaration of Independence	156
Author Contribution Statements	157

Summary

Introduction

'Natural selection is a mechanism for generating an exceedingly high degree of improbability'

- Ronald A. Fisher.

In 1859, Charles Darwin published his masterpiece *On the Origin of Species*, in which he proposed his theory of evolution by natural selection [1]. It remains the only known mechanism by which adaptation occurs. Despite the centrality of heritable variation to his theory, Darwin was unaware of the underlying molecular mechanisms. It was not until the fusion of Gregor Mendel's concepts of genetic inheritance with Darwin's theory of natural selection in the early 20th century, known as the modern synthesis, that a more complete picture of the adaptive process became possible. The result was the emergence of the field of population genetics.

Population genetics is based on the idea that natural selection and demographic events result in patterns of genetic variation in the DNA of a population. By studying these patterns we can reconstruct the evolutionary history of populations and species. The fate of genetic variation in a population is determined by the interplay of two forces; genetic drift and natural selection. Genetic drift changes allele frequencies due to random sampling while natural selection either raises the frequency of alleles that confer a fitness advantage (positive selection), maintains alleles at intermediate frequencies in cases of frequency-dependent selection or heterozygote advantage (balancing selection), or lowers the frequency of alleles that provide a fitness disadvantage (negative selection).

A key principle is the 'neutral theory of molecular evolution', which posits that the majority of genetic variation in the genome is selectively neutral, meaning it does not contribute to an organism's fitness [2]. Identifying the subset of variants that do influence fitness can provide insights into the biological basis for adaptations. A variety of tests have been developed to identify genomic loci with signatures of natural selection [3]. Those that detect signatures of positive selection are typically based on the *hard sweep*, in which a new mutation that confers a fitness advantage increases in frequency in the

Summary

population along with linked neutral variation until it reaches fixation [4]. This results in distinct localised patterns of genomic variation compared to background levels of neutral variation along the genome. Although other models of positive selection exist [5], the hard sweep model has proven to be an effective way to identify signatures of selection in genomic data [6, 7, 8].

Our understanding of adaptation in real populations has historically been constrained by a lack of data. Recent advances in genome sequencing technology and decreasing costs have enabled researchers to sequence large numbers of complete human genomes [9], providing the opportunity to study genomic signatures of natural selection and to identify causative variants that contribute to adaptation. However, focusing on human genetic variation alone cannot provide a complete picture of our evolutionary history. Comparable studies in our closest living relatives, the great apes, have been lacking. Without such studies it is impossible to distinguish which adaptive processes are shared across the *Hominidae* (humans and great apes) and which are uniquely human.

Studying the genomes of other species can also inform us about key transitions in human history. The domestication of a handful of animal species in the early holocene had a transformative impact on the development of human societies [10, 11]. Human management, controlled breeding and selection for traits that made them more suited to human needs, resulted in biological changes that make domesticated animals profoundly different from their wild ancestors [12]. Genomic studies of single species have provided insights into when and where domestication events occurred and identified loci that were positively selected by humans [13, 14]. However, a comparative genomic approach to examine whether the human induced changes are fundamentally similar or different in each species has been absent.

Research Purpose

I present three studies that use comparative population genomics to further our understanding of processes relevant to human evolution. In chapter 1, I investigate the genetic changes that were selected during the earliest stages of dog domestication. Dogs are thought to be the earliest animal species to be domesticated by humans [15], yet the genetic variants that underlie the phenotypes which were first selected by humans remain largely unknown. In chapter 2, I explore whether the phenotypic similarities observed across domesticated species have a shared genetic basis. I generate

Summary

genomic data from experimentally domesticated lines of rat and mink and analyse them in combination with population genomic data from seven pairs of historically domesticated species and their wild sister species. In chapter 3, I explore adaptation in our closest living relatives, the great apes. I analyse whole-genome population data from all the major *Hominidae* (humans and great apes) lineages. I investigate the impact of variation in population size on the efficacy of natural selection among these closely related lineages. I also identify signatures of positive, negative and balancing selection across the lineages. The purpose is to advance our understanding of human evolution by providing a more complete evolutionary context.

Chapter 1 - Identification of genomic variants putatively targeted by selection during dog domestication

The domestication of a few animal species in the early holocene had a transformative effect on the development of human societies. In their various forms domestic animals have provided food, labour, transport, raw materials, security and companionship for human groups. The productivity increases provided by domestic animals were essential for the development of complex hierarchical societies [10]. Furthermore, without domestic animals it is inconceivable that the human population would have been able to grow to its current size. Domestic animals have also been key models for understanding evolution. Studying the variation in domesticates due to artificial selection played a central role in inspiring Darwin's to formulate his theory of evolution by natural selection [16]. The Dog (*Canis lupus familiaris*) in particular holds a special place among domestic animals. It is considered to be the first animal domesticated by humans, with genetic and archaeological evidence suggesting this process started approximately 11-16 kya [17, 18].

Here, I present a comparative analysis of a global sample of dog and wolf genomes, to identify regions with signatures of selection and to identify the putatively causal variants that may have been involved in the earliest stages of dog domestication. I carried out a scan to identify putatively selection regions based on high levels of differentiation between dogs and wolves using windowed F_{st} . I found 18 regions with strong signatures of population divergence between dogs and wolves. As I use an outlier based approach to identify signatures of positive selection it is possible that the regions I detected are false-positives due to neutral demographic processes such as population bottlenecks, which can potentially produce similar patterns of genetic variation. To explore this I performed coalescent

Summary

simulations using a model of the demographic history of dogs and wolves [17]. I found that the putatively selected regions show higher divergence than any of the neutrally simulated regions, suggesting that their high levels of divergence cannot be explained by neutral evolution.

I identified all single nucleotide positions that are highly differentiated between dogs and wolves ($F_{st} \geq 0.75$) and that are predicted to be potentially functional. I performed a gene ontology enrichment analysis to gain insight into the 848 genes with such variants. I found that only the 'adrenaline and noradrenaline biosynthesis pathway' showed significant enrichment, suggesting that this pathway, known for its involvement in mediating the fight-or-flight response, may have been targeted by selection during early dog domestication. In addition, I identified 11 genes with putatively functional variants that are fixed for alternative alleles between dogs and wolves, three of which are implicated in neuronal development. The genes and pathways identified in this study provide new insights into the biological changes involved in the early stages of dog domestication.

Chapter 2 - Genes and pathways selected during animal domestication

Domestic animal species share a suite of morphological and behavioral traits. In comparison to their wild progenitors they tend to have more patches of white fur, smaller skulls, floppier ears and be more tame towards humans [19, 20]. This suite of traits is known as the 'domestication syndrome'. Since Darwin, evolutionary biologists have wondered what could explain these similarities. While there have been several genomic studies of domestic animals, these have typically been limited to single species without attempting to look for commonalities in the domestication process across species. While various hypotheses have been proposed [19, 20], we still lack a clear understanding of the biological changes involved in the animal domestication process. It remains unknown to what extent the phenotypic changes associated with domestication are due to direct selection for each trait, relaxation of selective pressures, genetic drift or correlated effects of selection for a single or few traits.

As the majority of domestic animals were domesticated thousands of years ago these questions have proved difficult to investigate in a quantitative manner. A remarkable exception are the experimental domestication experiments of the late Academician Dmitry Belyaev. He hypothesised that selection for tame behavior at the start of the domestication process had pleiotropic effects that explained the

Summary

phenotypic similarities observed across species. To test this, he experimentally domesticated wild-derived populations of foxes, rats and mink by selecting them solely for reduced fear towards humans [21]. In all three species this selection for tame behavior resulted in correlated phenotypic changes that echo many aspects of the domestication syndrome [22, 23, 24]. However, the underlying genetic changes, and whether they reflect changes that occurred during historical cases of animal domestication, remain unknown.

Here, I present comparative genomic analyses of the experimentally domesticated lines of rat (*Rattus norvegicus*) and mink (*Neovison vison*), to identify the genetic changes that underlie their phenotypic changes and to assess their relevance for understanding the domestication process. I developed a method to identify putatively selected genes that controls for biases due to gene length. I found that the rat and mink lines share a significant excess of putatively selected genes, suggesting a partially convergent genetic response to selection. To test whether the genetic changes in these lines are relevant models for understanding the domestication process I analysed whole-genome population data from seven pairs of historically domesticated animals and their wild sister species using the same approach. While I found no single 'domestication gene' selected in all species I identified six biological pathways appear to have been recurrently affected by the domestication process in all nine domesticated species. These results suggest that although the precise genetic changes selected during the domestication process vary between species, there are biological pathways that appear to always be involved.

Chapter 3 - Natural Selection in the Great Apes

A central challenge in evolutionary anthropology is to understand the adaptive genetic changes that led to the emergence of modern humans. To understand what makes humans unique it is necessary to compare ourselves to our closest living relatives, the great apes. While there have been extensive studies of signatures of natural selection in humans [3, 6], and some in the great apes [25, 26, 27, 28], no study has comprehensively investigated the signatures of natural selection across the *Hominidae*. Furthermore, most previous studies of adaptation in the great apes have relied on the analysis of single genomes, limiting the inferences that can be made in the absence of polymorphism data from larger samples.

Summary

Here, I present a global study of natural selection across the *Hominidae*, using genome-wide population data from all the major lineages. I applied multiple neutrality tests to create a comprehensive survey of positive, negative and balancing selection. Among regions with signatures of positive selection are several genes related to brain function and development, which may contribute to the advanced cognitive abilities of the *Hominidae*. I found that most signatures of positive and balancing selection are species specific. I determined that the differences in long-term effective population size between the *Hominidae* lineages have had a significant impact on the efficacy of both positive and negative selection. I also provide a global map of signatures of natural selection in the *Hominidae* as a public resource to aid future research.

Contribution to the field

The three studies presented here advance our understanding of key questions in evolutionary anthropology through the application of comparative genomics. In the first study, I used a global sample of dog and wolf genomes to identify genetic variants putatively involved in the early stages of dog domestication. This work advances our understanding of this process and suggests that early dog domestication involved changes in genes related to the fight-or-flight response. In the second study, I analysed genomes from experimentally and historically domesticated animal species. This work provides the first genomic evidence that some biological pathways are always affected during animal domestication. In the third study, I applied neutrality tests to genomes from population samples of all the major *Hominidae* lineages. This advances the field by providing the most comparative survey to date of positive, negative and balancing selection in humans and our closest living relatives. I also demonstrated the important role that effective population size has had on the efficacy of selection in these lineages, providing empirical support to theoretical predictions about the relationship between population size and adaptation.

Outlook

As sequencing costs decrease the increase in polymorphism data will enable researchers to make even more detailed investigations into genomic signatures of adaptation. The rise of functional genomics and genome-editing will provide methods to elucidate the molecular mechanisms by which

Summary

the variants identified in these studies contribute to adaptation. This will lead to new insights into human evolutionary history.

Zusammenfassung

Einführung

"Natürliche Selektion ist ein Mechanismus der ein außerordentlich hohes Maß an Unwahrscheinlichkeit erzeugt."

- Ronald A. Fisher

Im Jahre 1859 veröffentlichte Charles Darwin sein Meisterwerk *“Über die Entstehung der Arten”*, in dem er seine Evolutionstheorie durch natürliche Selektion aufstellte [1]. Sie bleibt der einzige bekannte Mechanismus, durch den Anpassung stattfindet. Obwohl die vererbte Variation in Darwins Theorie von zentraler Bedeutung war, waren ihm die zugrunde liegenden molekularen Mechanismen nicht bewusst. Erst durch die Verschmelzung mit Gregor Mendels Vorstellungen der genetischen Vererbung mit Darwins Theorie der natürlichen Selektion im frühen 20. Jahrhundert, die als moderne Synthese bezeichnet wird, wurde ein vollständigeres Bild der adaptiven Prozesse möglich. Das Ergebnis war die Entstehung des Feldes der Populationsgenetik.

Populationsgenetik basiert auf der Idee, dass natürliche Selektion und demographische Ereignisse zu Mustern von genetischer Variation in der DNA einer Population führen. Durch das Studium dieser Muster können wir die Evolutionsgeschichte von Populationen und Arten rekonstruieren. Das Schicksal der genetischen Variation in einer Population wird durch das Zusammenspiel zweier Kräfte bestimmt; Genetische Drift und natürliche Selektion. Die genetische Drift ändert die Allelhäufigkeiten aufgrund zufälliger Stichprobenentnahme, während die natürliche Selektion entweder die Häufigkeit von Allelen, die einen Fortpflanzungsvorteil verleihen, erhöht (positive Selektion), oder die Häufigkeit von Allelen, die einen Fortpflanzungsnachteil bedeuten, senkt (negative Selektion). Zusätzlich kann im Falle von frequenzabhängiger Selektion sowie bei vorteilhaften Heterozygoten die Häufigkeit von Allelen in einem intermediären Bereich stabilisieren werden (balancierende Selektion).

Ein Schlüsselprinzip ist die "Neutrale Theorie der molekularen Evolution", die besagt, dass die Mehrheit der genetischen Variation im Genom selektiv neutral ist, was bedeutet, dass sie nicht zur

Zusammenfassung

Fitness eines Organismus beiträgt [2]. Die Identifizierung der Teilmenge an Varianten, die die Fitness beeinflussen, kann Einblicke in die biologische Grundlage für Anpassungen liefern. Es wurden eine Vielzahl von Tests entwickelt, um genomische Loci mit Signaturen natürlicher Selektion zu identifizieren [3]. Diejenigen, die Signaturen positiver Selektion erkennen können, basieren typischerweise auf dem *hard sweep*, bei dem eine neue Mutation, die einen Fitnessvorteil verleiht, in der Population zusammen mit der mit ihr verbundenen neutralen Variation zunimmt, bis sie die Fixierung erreicht [4]. Dies führt zu deutlichen ortsgebundenen Mustern der genomischen Variation im Vergleich zu dem Grundniveau der neutralen Variation entlang des Genoms. Obwohl andere Modelle der positiven Selektion existieren [5], hat sich das *hard sweep*-Modell als wirksame Methode zur Identifizierung von Signaturen der Selektion in genomischen Daten erwiesen [6, 7, 8].

Unser Verständnis der Anpassung in realen Populationen wurde in der Vergangenheit durch einen Mangel an Daten eingeschränkt. Die jüngsten Fortschritte in der Genom-Sequenzierungstechnologie und die sinkenden Kosten haben es den Forschern ermöglicht, eine große Anzahl vollständiger menschlicher Genome zu erfassen [9], die die Möglichkeit bieten, genomische Signaturen der natürlichen Selektion zu untersuchen und ursächliche Varianten zu identifizieren, die zur Anpassung beitragen. Allerdings kann die Fokussierung auf die menschliche genetische Variation allein kein vollständiges Bild von unserer Evolutionsgeschichte geben. Vergleichbare Studien in unseren engsten lebenden Verwandten, den Menschenaffen, fehlten. Ohne solche Studien ist es unmöglich zu unterscheiden, welche adaptiven Prozesse von den Hominidae (Menschen und Menschenaffen) geteilt werden und welche eindeutig dem Menschen zuzuordnen sind.

Das Studium der Genome anderer Arten kann uns auch über Schlüsselübergänge in der menschlichen Geschichte informieren. Die Domestizierung einer Handvoll Tierarten im frühen Holozän hatte einen tiefgreifenden Einfluss auf die Entwicklung der menschlichen Gesellschaften [10, 11]. Die menschliche Bewirtschaftung, die kontrollierte Zucht und die Selektion für Merkmale, die sie besser auf die menschlichen Bedürfnisse abstimmten, führten zu biologischen Veränderungen, durch die sich die domestizierte Tiere zutiefst von ihren wilden Vorfahren unterscheiden [12]. Genomische Studien von einzelnen Arten haben Einblicke gegeben, wann und wo Domestizierungsereignisse auftraten, und Loci identifiziert, die von Menschen positiv selektiert worden sind [13, 14]. Jedoch fehlt ein vergleichender genomischer Ansatz, um zu untersuchen, ob die vom Menschen induzierten Veränderungen in jeder Spezies grundsätzlich ähnlich oder unterschiedlich sind.

Forschungszweck

Ich präsentiere drei Studien, die vergleichende Populationsgenomik einsetzen, um unser Verständnis der für die menschliche Evolution relevanten Prozesse zu fördern. In Kapitel 1 untersuche ich die genetischen Veränderungen, die während der frühesten Stadien der Hundedomestizierung ausgewählt wurden. Hunde sollen die älteste Tierart sein, die von Menschen domestiziert wurden [15], doch die genetischen Varianten, die den Phänotypen zugrunde liegen, die zuerst von Menschen ausgewählt wurden, sind weitgehend unbekannt. In Kapitel 2 erforsche ich, ob die phänotypischen Ähnlichkeiten, die bei domestizierten Arten beobachtet wurden, eine gemeinsame genetische Basis haben. Ich generiere genomische Daten aus experimentell domestizierten Linien von Ratten und Nerzen und analysiere sie in Kombination mit Populationsgenomdaten von sieben Paaren historisch domestizierter Spezien und ihrer Wildschwesterarten. In Kapitel 3 erforsche ich die Anpassung in unseren engsten lebenden Verwandten, den Menschenaffen. Ich analysiere vollständige Genom-Populationsdaten von allen großen Hominidae Abstammungslinien (Menschen und Menschenaffen). Ich untersuche die Auswirkungen der Variation der Populationsgröße auf die Wirksamkeit der natürlichen Selektion zwischen diesen nah verwandten Linien. Ich identifiziere auch Signaturen positiver, negativer und balancierender Selektion über die Abstammungslinien hinweg. Der Zweck ist, unser Verständnis der menschlichen Evolution voranzutreiben, indem wir einen vollständigeren evolutionären Kontext bereitstellen.

Kapitel 1 - Identifizierung von genomischen Varianten, die vermeintlich gezielt während der Domestizierung von Hunden selektiert wurden

Die Domestizierung einiger Tierarten im frühen Holozän hatte eine veränderte Wirkung auf die Entwicklung der menschlichen Gesellschaften. In ihren verschiedenen Formen haben die Haustiere Nahrung, Arbeit, Transport, Rohstoffe, Sicherheit und Gesellschaft für menschliche Gruppen zur Verfügung gestellt. Die Produktivitätssteigerungen von Haustieren waren für die Entwicklung komplexer hierarchischer Gesellschaften essentiell [10]. Darüber hinaus ist es ohne Haustiere nicht vorstellbar, dass die menschliche Bevölkerung in der Lage gewesen wäre, auf ihre aktuelle Größe zu wachsen. Domestizierte Tiere sind auch Schlüsselmodelle für das Verständnis der Evolution. Das Studium der Variation in Haustieren, die durch künstliche Selektion hervorgerufen wurde, spielte eine

Zusammenfassung

zentrale Rolle für die Inspiration Darwins, seine Evolutionstheorie auf Grund von natürlicher Selektion zu formulieren [16]. Der Hund (*Canis lupus familiaris*) hält einen besonderen Platz unter den Haustieren: er gilt als das erste von Menschen domestizierte Tier. Genetische und archäologische Hinweisen deuten darauf hin, dass dieser Prozess etwa vor 11.000 – 16.000 Jahren begann [17, 18].

Hier präsentiere ich eine vergleichende Analyse einer globalen Stichprobe von Hunde- und Wolfgenomen, um Regionen im Genom mit Signaturen der Selektion zu identifizieren und die vermeintlich kausalen Varianten zu identifizieren, die an den frühesten Stadien der Hundedomestikation beteiligt gewesen sein könnten. Ich habe einen Scan durchgeführt, um mutmaßlich selektierte Regionen basierend auf der Berechnung der Unterschiede in F_{ST} für einzelne Fenster zwischen Hunden und Wölfen zu identifizieren. Ich habe 18 Regionen mit deutlichen Signaturen der Divergenz zwischen Hunde- und Wolfpopulationen gefunden. Da ich einen auf statistische Ausreißer basierenden Ansatz zur Identifizierung von Signaturen der positiven Selektion nutze, ist es möglich, dass die von mir erkannten Regionen aufgrund von neutralen demographischen Prozessen, wie etwa Bevölkerungsengpässen, ähnliche Muster der genetischen Variation haben können und somit falsch positiv sind. Um dies zu erforschen, habe ich Koaleszenzsimulationen basierend auf einem Modell der demographischen Geschichte von Hunden und Wölfen durchgeführt [17]. Ich habe herausgefunden, dass die vermeintlich selektierten Regionen eine höhere Divergenz aufweisen als irgendwelche der als neutral simulierten Regionen, was darauf hindeutet, dass ihre hohe Divergenz nicht durch eine neutrale Evolution erklärt werden kann.

Ich habe alle einzelnen Nukleotidpositionen identifiziert, die zwischen Hunden und Wölfen stark differenziert ($F_{ST} > = 0,75$) und potentiell funktional sind. Ich habe eine Gen-Ontologie-Anreicherungsanalyse durchgeführt, um Einblick in die 848 Gene mit solchen Varianten zu gewinnen. Ich habe fest gestellt, dass nur der "Adrenalin- und Noradrenalin-Biosyntheseweg" eine signifikante Anreicherung zeigte, was darauf hindeutet, dass dieser Weg, der für seine Beteiligung an der Vermittlung der Kampf- oder Fluchtreaktion bekannt ist, während der frühen Hundedomestikation gezielt selektiert worden sein könnte. Darüber hinaus habe ich 11 Gene mit putativ funktionellen Varianten identifiziert, für die das alternative Allele bezogen auf Hunde und Wölfe fixiert ist; von diesen sind drei in neuronaler Entwicklung verwickelt. Die Gene und Stoffwechselwege, die in dieser Studie identifiziert worden sind, geben neue Einblicke in die biologischen Veränderungen in den frühen Stadien der Hundedomestikation.

Kapitel 2 - Gene und Stoffwechselwege, die während der Tierdomestikation selektiert wurden

Domestizierte Tierarten teilen sich eine Reihe von morphologischen und Verhaltensmerkmalen. Im Vergleich zu ihren wilden Vorläufern neigen sie dazu, mehr weiße Flecken im Pelz, kleinere Schädel, und Schlappohren zu haben und zahmer gegenüber Menschen zu sein [19, 20]. Diese Gruppe von Merkmalen wird als "Domestizierungssyndrom" bezeichnet. Seit Darwin haben sich evolutionäre Biologen gefragt, was diese Gemeinsamkeiten erklären könnte. Während es mehrere genomische Studien von Haustieren gab, waren diese typischerweise auf einzelne Arten beschränkt, ohne zu versuchen, nach Gemeinsamkeiten im Domestizierungsprozess über Artgrenzen hinweg zu suchen. Während verschiedene Hypothesen vorgeschlagen wurden [19, 20], fehlt es uns noch an einem klaren Verständnis der biologischen Veränderungen, die an der Tierdomestikation beteiligt sind. Wir wissen noch nicht, inwieweit die phänotypischen Veränderungen, die mit der Domestikation verbunden sind, sich auf die direkte Selektion für jedes einzelne Merkmal, die Lockerung von selektiven Drücken, den Gendrift oder die damit korrelierten Effekte der Selektion für ein oder mehrere Merkmale zurückführen lassen.

Da die Mehrheit der Haustiere vor Tausenden von Jahren domestiziert wurde, hat sich die quantitative Untersuchung dieser Fragen als schwierig erwiesen. Eine bemerkenswerte Ausnahme sind die experimentellen Domestikationsexperimente des verstorbenen Akademikers Dmitri Beljajew. Er vermutete, dass die Selektion für das zahme Verhalten zu Beginn des Domestizierungsprozesses pleiotrope Effekte hatte, die die phänotypischen Ähnlichkeiten, die bei den Arten beobachtet wurden, erklären. Um dies zu testen, hat er experimentell aus der Wildnis bezogene Populationen von Füchse, Ratten und Nerze domestiziert, indem er sie nur bezüglich einer verminderte Angst gegenüber Menschen ausgewählt hat [21]. In allen drei Arten führte diese Selektion für das zahme Verhalten zu korrelierten phänotypischen Veränderungen, die viele Aspekte des Domestizierungssyndroms widerspiegeln [22, 23, 24]. Allerdings bleiben die zugrunde liegenden genetischen Veränderungen und ob sie Veränderungen widerspiegeln, die auch während der historischen Fälle von Tierdomestikation auftraten, unbeantwortet.

Zusammenfassung

Hier präsentiere ich vergleichende genomische Analysen der experimentell domestizierten Rattenlinien (*Rattus norvegicus*) und Nerz (*Neovison vison*), um die genetischen Veränderungen zu identifizieren, die den phänotypischen Veränderungen zugrunde liegen und deren Relevanz für das Verständnis des Domestizierungsprozesses zu beurteilen. Ich habe eine Methode zur Identifizierung vermeintlich selektierter Gene entwickelt, die systematische Fehler aufgrund der Genlänge berücksichtigt. Ich habe herausgefunden, dass die Ratten- und Nerz-Linien sich einen signifikanten Überschuss an vermeintlich selektierten Genen teilen, was auf eine teilweise konvergente genetische Antwort auf die Selektion hindeutet. Um zu testen, ob die genetischen Veränderungen in diesen Linien als relevante Modelle für das Verständnis des Domestizierungsprozesses dienen können, habe ich die Gesamtgenom-Populationsdaten aus sieben Paaren historisch domestizierter Tiere und ihrer Wildschwesterspezies mit demselben Ansatz analysiert. Während ich kein einziges "Domestizierungsgen" in allen Arten finden konnte, habe ich sechs biologische Stoffwechselwege entdeckt, die immer wieder in allen neun domestizierten Arten durch die Domestikation beeinflusst worden sind. Diese Ergebnisse deuten darauf hin, dass es, obwohl die genauen genetischen Veränderungen, die während des Domestikationsprozesses ausgewählt werden, zwischen den Arten variieren, biologische Wege gibt, die immer beteiligt zu sein scheinen.

Kapitel 3 - Natürliche Selektion in Menschenaffen

Eine zentrale Herausforderung in der evolutionären Anthropologie ist es, die adaptiven genetischen Veränderungen zu verstehen, die zum Entstehen moderner Menschen führten. Um zu verstehen, was den Menschen einzigartig macht, ist es notwendig, uns mit unseren engsten lebenden Verwandten, den Menschenaffen, zu vergleichen. Während es umfangreiche Studien über Signaturen der natürlichen Selektion bei Menschen [3, 6] und einige in den Menschenaffen gibt [25, 26, 27, 28], hat keine Studie die Signaturen der natürlichen Selektion über die Hominidae hinweg umfassend untersucht. Darüber hinaus haben sich die meisten früheren Studien über die Anpassung in den Menschenaffen auf die Analyse einzelner Genome gestützt und limitieren somit die möglichen Schlussfolgerungen, die in Abwesenheit von Polymorphismusdaten aus größeren Probenmengen gemacht werden können.

Hier präsentiere ich eine globale Studie über die natürliche Selektion über die Hominidae hinweg, wobei genomweite Populationsdaten aus allen Hauptabstammungslinien verwendet werden. Ich habe

Zusammenfassung

mehrere Neutralitätstests angewendet, um eine umfassende Übersicht über die positive, negative und balancierende Selektion zu erstellen. Unter den Regionen mit Signaturen der positiven Selektion sind mehrere Gene im Zusammenhang mit Hirnfunktion und Entwicklung zu finden, die zu den fortgeschrittenen kognitiven Fähigkeiten der Hominidae beigetragen haben können. Ich habe festgestellt, dass die meisten Signaturen der positiven und balancierenden Selektion artspezifisch sind. Ich habe festgestellt, dass die Unterschiede in der langfristig wirksamen Populationsgröße zwischen den Hominidae-Linien einen signifikanten Einfluss auf die Wirksamkeit sowohl der positiven als auch der negativen Selektion hatten. Ich stelle auch eine globale Karte der Signaturen der natürlichen Selektion in den Hominidae für die Öffentlichkeit zur Verfügung, um der zukünftige Forschung zu helfen.

Beitrag zum Forschungsfeld

Die drei hier vorgestellten Studien bringen unser Verständnis der Schlüsselfragen in der evolutionären Anthropologie durch die Anwendung der vergleichenden Genomik voran. In der ersten Studie habe ich eine globale Stichprobe von Hunde- und Wolfgenomen verwendet, um genetische Varianten zu identifizieren, die vermutlich in die frühen Stadien der Hundedomestikation involviert sind. Diese Arbeit erweitert unser Verständnis für diesen Prozess und deutet darauf hin, dass die frühe Hundedomestizierung Veränderungen in Genen im Zusammenhang mit der Kampf-oder-Flucht-Reaktion hervorgerufen hat. In der zweiten Studie habe ich Genome aus experimentell und historisch domestizierten Tierarten analysiert. Diese Arbeit liefert die ersten genomischen Beweise, dass einige biologische Wege immer während der Tierdomestikation beeinflusst werden. In der dritten Studie habe ich Neutralitätstests auf Genome aus Populationsstichproben aller Hauptlinien der Hominidae angewendet. Dies bringt das Forschungsfeld voran, indem es die vergleichenste Untersuchung zur positiven, negativen und balancierenden Selektion in Menschen und unseren engsten lebenden Verwandten zur Verfügung stellt. Ich habe auch die wichtige Rolle aufgezeigt, die die effektive Populationsgröße auf die Wirksamkeit der Selektion in diesen Linien hat, und biete Unterstützung aufgrund empirischer Daten für die theoretischer Vorhersagen über die Beziehung zwischen Bevölkerungsgröße und Anpassung.

Ausblick

Zusammenfassung

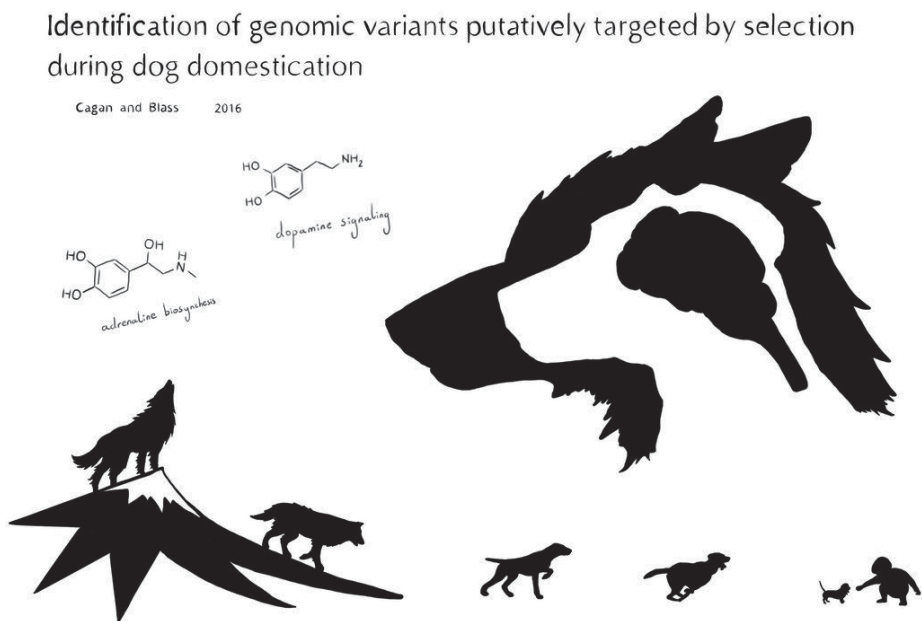
Da die Sequenzierungskosten sich verringern, wird der Anstieg der Polymorphismus-Daten Forschern erlauben, noch detailliertere Untersuchungen zu genomischen Signaturen der Anpassung durchzuführen. Der Aufstieg der funktionellen Genomik und des Genom-Editierens wird Methoden zur Aufklärung der molekularen Mechanismen liefern, durch die die in diesen Studien identifizierten Varianten zur Anpassung beitragen haben. Dies wird zu neuen Einsichten in die menschliche Evolutionsgeschichte führen.

Chapter 1

Identification of genomic variants putatively targeted by selection during dog domestication

Published in
BMC Evolutionary Biology, (2016).

by
Alex Cagan, Torsten Blass



RESEARCH ARTICLE

Open Access



Identification of genomic variants putatively targeted by selection during dog domestication

Alex Cagan*  and Torsten Blass

Abstract

Background: Dogs [*Canis lupus familiaris*] were the first animal species to be domesticated and continue to occupy an important place in human societies. Recent studies have begun to reveal when and where dog domestication occurred. While much progress has been made in identifying the genetic basis of phenotypic differences between dog breeds we still know relatively little about the genetic changes underlying the phenotypes that differentiate all dogs from their wild progenitors, wolves [*Canis lupus*]. In particular, dogs generally show reduced aggression and fear towards humans compared to wolves. Therefore, selection for tameness was likely a necessary prerequisite for dog domestication. With the increasing availability of whole-genome sequence data it is possible to try and directly identify the genetic variants contributing to the phenotypic differences between dogs and wolves.

Results: We analyse the largest available database of genome-wide polymorphism data in a global sample of dogs 69 and wolves 7. We perform a scan to identify regions of the genome that are highly differentiated between dogs and wolves. We identify putatively functional genomic variants that are segregating or at high frequency [$> = 0.75 F_{st}$] for alternative alleles between dogs and wolves. A biological pathways analysis of the genes containing these variants suggests that there has been selection on the 'adrenaline and noradrenaline biosynthesis pathway', well known for its involvement in the fight-or-flight response. We identify 11 genes with putatively functional variants fixed for alternative alleles between dogs and wolves. The segregating variants in these genes are strong candidates for having been targets of selection during early dog domestication.

Conclusions: We present the first genome-wide analysis of the different categories of putatively functional variants that are fixed or segregating at high frequency between a global sampling of dogs and wolves. We find evidence that selection has been strongest around non-synonymous variants. Strong selection in the initial stages of dog domestication appears to have occurred on multiple genes involved in the fight-or-flight response, particularly in the catecholamine synthesis pathway. Different alleles in some of these genes have been associated with behavioral differences between modern dog breeds, suggesting an important role for this pathway at multiple stages in the domestication process.

Keywords: Genomics, Domestication, Artificial selection, Natural selection, Behavioural genomics

* Correspondence: alexander_cagan@eva.mpg.de
Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany



© 2016 Cagan and Blass. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Background

Dogs [*Canis lupus familiaris*] are considered the first animal species to be domesticated by humans. Genetic and archaeological evidence suggests that this process began approximately 11-16kya [1, 2]. Dogs and their closest living relatives, wolves [*Canis lupus*] differ in a variety of phenotypic traits, despite only differing in ~0.047 % of nuclear coding-DNA sequence [3]. Particular attention has been given to their behavioral differences, with dogs showing a greater ability to read human communicative behaviour [4]. When and how these new cognitive abilities emerged remains unclear. It has been suggested that rather than selection for these specific behaviors it was selection for tameness, a reduction in fear and aggression towards humans, that permitted the expression of these latent abilities, which are inhibited in wolves by their fear response [5, 6].

Unlike the majority of domestic species, which were primarily selected for production related traits, dogs were typically selected for their behaviors [7]. Modern breeds are the result of human mediated selection for an incredibly wide-range of behaviors, including guarding, herding and pointing [8]. Pioneering early work on breed crosses demonstrated a genetic basis to some of these behavioral differences between breeds [9, 10]. Since then much work has been done to identify the genetic basis of phenotypic differences between dog breeds. The great phenotypic diversity and population structure between modern dog breeds has proven to be a powerful model for elucidating the genetic basis of breed-specific traits [3]. Studies have utilized a variety of approaches including trait mapping [11, 12] selection scans [12, 13] and candidate gene driven approaches [14, 15].

There has been much success in identifying genetic variants underlying morphological traits, which often have a relatively simple mono-allelic genetic architecture [12, 16, 17]. Identifying the genetic basis of behavioral traits, which are typically assumed to have a more complex genetic architecture, has proven to be a more challenging endeavor [3]. Nevertheless, canine behavioral genetics is a rapidly moving field and several studies have made progress in uncovering the genetic variants associated with behavioral differences between breeds [8, 18, 19].

One behaviour of particular interest is aggression, given that selection for reduced aggression towards humans was likely a prerequisite for domestication [20]. Indeed, dogs generally show reduced fear and aggression towards humans compared to wolves [21]. Candidate gene approaches have identified significant allele frequency differences that correlate with levels of aggression related behaviour within or between dog breeds in genes that have previously been associated with

aggression in humans. Examples include monoamine oxidase B [*MAOB*] [22], the dopamine D4 receptor [*DRD4*] [23], the dopamine transporter [*SLC6A3*] [24], tyrosine hydroxylase [*TH*] [25] and dopamine beta-hydroxylase [*DHB*] [25]. One study tested 62 SNPs occurring within or close to 16 neurotransmitter-related genes for allelic associations with aggression [26]. Although multiple risk or protective haplotypes for aggression were identified no single haplotype was in complete association with the phenotypes recorded, supporting the view that aggressive behaviour in dogs has a complex genetic basis. Taken together these results suggest that selection for behavioral traits related to aggression in dogs has targeted a variety of pathways, particularly those involving the synthesis, transport and degradation of neurotransmitters such as dopamine.

Despite this progress the genetic changes underlying reduced fear and aggression in dogs relative to wolves remain unknown. It is not necessarily the case that the genes associated with breed-specific behaviors are the same ones that were selected during the early domestication process. Indeed, despite the success of breed mapping approaches, their dependence on inter-breed variation makes them unsuitable to identify genetic variants selected for during the early domestication process that are shared by all dog breeds. While the findings of studies that focus on intra-breed variation may not be generalizable across breeds. As a result we know less about the genetic basis of the phenotypic changes that occurred during the early stages of dog domestication and differentiate all dogs from their wild progenitors than we do about differences between modern dog breeds.

Identifying the genetic changes that occurred early in the domestication process thus necessitates additional approaches beyond comparisons between breeds. Gene expression studies have identified sets of genes that are differentially expressed in the brains of dogs and wolves [27, 28] and between aggressive and non-aggressive dog breeds [29], however whether these contribute to behavioral differences remains unknown. Previous work using scans for selection in genomic data from dogs and wolves has identified genomic regions that may have been targeted by selection during early dog domestication [30–34]. In most cases the putative causative genomic variants underlying these selection signals remain to be identified. In most cases the putative causative genomic variants underlying these selection signals remain to be identified. One of the few cases where the causative variant has been identified is the gene *AMY2B*. Axelsson et al. [32] found that modern dogs have increased copy numbers of the pancreatic amylase gene *AMY2B* compared to wolves, potentially an adaptation

to a starch rich diet associated with human cohabitation. Although a later study found that this variation is polymorphic and does not represent a truly fixed genetic difference between dogs and wolves [1].

Thus far the putatively functional variants that are fixed or segregating at high frequency between dogs and wolves have not been systematically characterized. One exception is the study of Li et al. [35], in which non-synonymous variants segregating for alternative alleles between dogs and wolves were identified. However, this study was limited by a relatively small sample size [three wolves and five dogs], meaning that many of the sites they identified may not be truly segregating between all dogs and wolves. Furthermore, they only considered non-synonymous variants as putatively functional. Identifying and studying the properties of a wide range of putatively functional variants is of interest because they are expected to include the alleles that were selected during dog domestication and are responsible for the phenotypic differences between dogs and wolves. Furthermore, studies that rely solely on selection scans to identify adaptive loci are liable to false positives due to hitchhiking of neutral variants, particularly in populations that have experienced strong bottlenecks [36], such as domestic dogs [1]. Prioritising candidate regions that contain putatively functional variants is one way to increase the likelihood of identifying the true selective sweeps.

We analyzed variants that are fixed or segregating at high frequency between dogs and wolves. We identified these variants using DoGSD, the largest available dataset of whole-genome polymorphism data from dogs and wolves [37]. Of these variants we identify a subset as being putatively functional. We combine this information with a genomic scan for selection to identify regions of the genome that are highly diverged between dogs and wolves. We perform Gene Ontology analysis of genes with putatively functional variants segregating at high frequency between dogs and wolves. We find that putatively functional changes influencing genes involved in adrenaline biosynthesis appear to have been particularly targeted by selection during dog domestication. We find that selection during dog domestication appears to have been strongest around variants influencing protein structure. Furthermore, we identify 11 genes with putatively functional variants that appear to be fixed for alternative alleles between dogs and wolves. These changes are of particular interest because they may be the genetic variants responsible for the phenotypic differences between all dogs and all wolves that may have been selected during dog domestication.

Results and discussion

Scan for selection

To identify genomic regions that may contain variants that were selected during dog domestication we identified regions that were highly diverged between dogs and wolves by calculating the mean F_{st} between dogs and wolves in 500kb windows along the genome. Although previous studies have performed window-based scans for signatures of selection in dogs and wolves [30, 32], none have been performed on such a large sample of either species using whole-genome data. Following Axelsson et al. [32] we Z transform our F_{st} scores and consider regions scores that fall at least five standard deviations from the mean ($Z(F_{st})$) as putatively selected (Fig. 1).

Mean levels of divergence are higher on the X chromosome (X chromosome mean F_{st} = 0.21 compared to 0.14 for autosomes). This is usually attributed to the smaller effective population size of the X chromosome due to its mode of transmission [38]. However, it is also possible that this signal is partially the result of artificial selection during domestication having occurred disproportionately on the X chromosome. As males are hemizygous for X-linked traits this may have provided humans with an opportunity to easily identify and select recessive alleles on the X chromosome. As the penetrance of any given genetic variant in a population is dependent on its allele frequency and its mode of dominance, regardless of underlying demographic history, we use the same threshold to identify putatively selected regions on the X chromosome and the autosomes. We acknowledge that this may result in a higher false positive rate on the X chromosome. When the X chromosome is considered independently no regions on the X chromosome fall over five standard deviations above the mean F_{st} score. Nevertheless, as mentioned above, the X chromosome may contain functional variation contributing to dog domestication and we do not want to miss true positives through an overly stringent cut-off. Therefore, following Li et al. [35], we include the X chromosome in our selection analyses.

Using these criteria we identify 18 regions with strong signatures of population divergence between dogs and wolves (Table 1). As expected from the higher levels of mean divergence on the X chromosome, 13 of these regions are on this chromosome. 14 of these 18 regions contain genes which are candidates for being targets of selection. We identify many regions previously found to be under selection in dogs, including a region on chromosome 1 containing *MBP*, which encodes myelin basic protein, and a region on chromosome 16 which contains *MGAM*, involved in starch metabolism [32].

The selection scan was performed on a larger and more geographically diverse dataset than previous scans for selection comparing dogs and wolves [30–35]. We

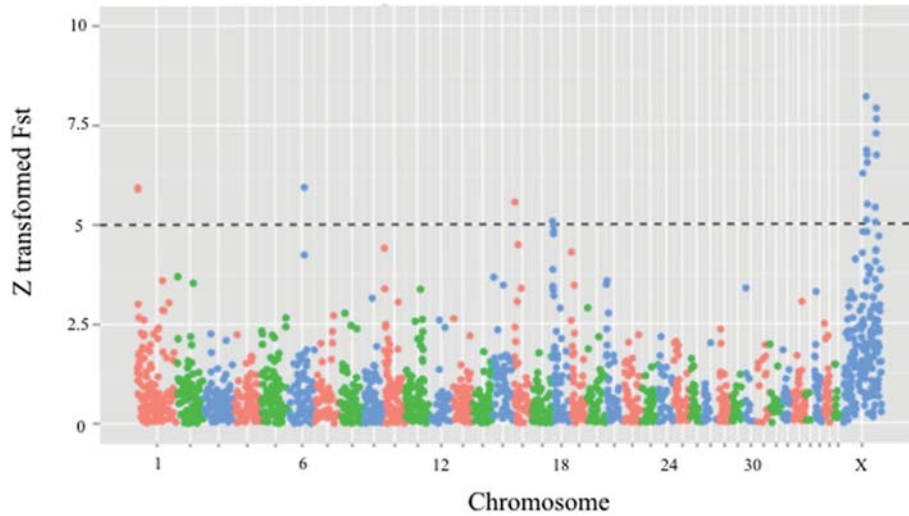


Fig. 1 Genome-wide scan for selective sweeps. Z-transformed mean Fst calculated in 500kb genomic windows across the autosomes and X chromosome between dogs and wolves. Each point represents a 500kb window. A dashed horizontal line represents our threshold for identifying putatively selected regions (>5 Z(Fst)). 18 windows exceed this threshold and are considered as putative selective sweeps

note that while our dataset was chosen to sample as broadly as possible from the worldwide distribution of dog and wolf populations that our dog sample is particularly enriched for German Shepherds [11], Tibetan Mastiffs [11] and indigenous dogs [39]. Therefore, the sweep signals that we detect may be shared only among these breeds and not truly reflect universal signatures of selection across dog breeds. Future studies with sampling from across a wider range of breeds will be

necessary to confirm whether these regions have elevated divergence between all dog and wolf populations.

To explore whether the elevated mean Fst in these regions could be explained by neutral evolutionary processes rather than selection we performed coalescent simulations for the autosomal genome based on a neutral model of the demographic history of these samples (Materials and Methods). We simulated 500kb haplotypes for all samples and calculated mean Fst between

Table 1 Genes in 500kb windows with Z transformed mean Fst scores five standard deviations above the mean

Chromosome	Window [bp]	Mean Fst	Genes in window
1	2500001–3000000	0.427497	SNORA70, GALR1 MBP, ZNF236
1	3000001–3500000	0.429086	U6, ZNF516
6	47000001–47500000	0.429446	RNPC3
16	7000001–7500000	0.411249	PRSS58, MGAM, TAS2R38, CLEC5A, PRSS37, U6, TAS2R5, TAS2R3, SSBP1, WEE2
18	500001–1000000	0.388113	
X	66000001–66500000	0.445258	ZNF711, POF1B, 7S
X	77000001–77500000	0.53831	TCEAL1, MORF4L2, GLRA4, TMEM31, PLP1, RAB9B, SLC25A53, 7SK,
X	77500001–78000000	0.388753	FAM199X, ESX1
X	78000001–78500000	0.473412	
X	79500001–80000000	0.46818	U6, TBG, MUM1L1
X	80000001–80500000	0.458387	U6, CXorf57, RNF128, RNF128, TBC1D8B, CLDN2, RIPPLY1, MORC4
X	80500001–81000000	0.407971	RBM41, NUP62CL, PIH1D3
X	105500001–106000000	0.404133	MOSPD1, ZNF75D, U6, ZNF449, DDX26B
X	106000001–106500000	0.386268	SLC9A6
X	108000001–108500000	0.493717	
X	108500001–109000000	0.524088	FGF13, cfa-mir-504
X	109000001–109500000	0.467526	
X	109500001–110000000	0.511206	F9, MCF2, U4, ATP11C, cfa-mir-505, CXorf66

the pooled dog and wolf haplotypes so that the results could be compared to our empirical data. The mean of the mean F_{st} scores across all simulations is slightly elevated, $F_{st} = 0.184$, compared to the mean F_{st} of our real data, $F_{st} = 0.144$, or when excluding the X chromosome, $F_{st} = 0.140$. Despite this elevated mean F_{st} , we never observe simulated 500kb regions with mean F_{st} scores as high as our putatively selected regions (Additional file 1: Figure S3). The highest mean F_{st} score from the simulations is 0.31, while the lowest mean F_{st} score of the 18 putatively selected regions is 0.39. Therefore, the simulations suggest that the cut-off we use to detect putatively selected regions is conservative and the elevated mean F_{st} scores of these regions are unlikely to have been the result of purely neutral evolutionary forces.

Variants fixed for alternative alleles between dogs and wolves

As many of these putatively selected regions contain multiple genes the identification of the targets of selection is challenging. There may also be selected variants that are not surrounded by the signatures of a selective sweep. This could occur for a variety of reasons, including when selection occurs on standing genetic variation [40] and because strong population bottlenecks reduce our ability to detect signatures of selection over neutrality [36]. Both these scenarios appear to have occurred during the process of dog domestication [1].

To try and identify the targets of selection in these putatively selected regions as well as selected variants not surrounded by signatures of selection we identified all single nucleotide positions that were fixed for alternative

alleles between dogs and wolves ($F_{st} = 1$). From this list of 2112 sites we used Ensembl's Variant Effect Predictor (VEP) to identify those which had putatively functional consequences [41] (Materials and Methods).

We identify only 11 genes with putatively functional positions that appear fixed for alternative alleles between dogs and wolves (Table 2). Eight of these fall within the selective sweep regions. Of the remaining four, three are in 500kb windows directly neighbouring the candidate selective sweep regions. The remaining gene, *RELT*, is in the ninth most highly diverged 500kb region between dogs and wolves on chromosome 21. Therefore, the majority of fixed putatively functional variants are found regions within highly diverged regions, suggesting that for dog domestication a hard sweep model may be appropriate for detecting selected variants. The relatively low N_e of the population ancestral to all dogs, estimated to be as low as 700–3,200 [1], combined with the high selection coefficients possible under artificial selection, may have increased the likelihood of hard sweeps relative to other non-domesticated species where selection has been studied, such as *Drosophila melanogaster* [42].

A previous study on dog domestication by Li et al. [35] identified 26 non-synonymous variants that were fixed for alternative alleles between dogs and wolves. Using our larger dataset we were able to further refine this list. Of the 26 non-synonymous variants they identified, only six appear as true substitutions between dogs and wolves in our analysis. Five of these six substitutions fall in two genes of unknown function on chromosome X (*ENSCAF00000018988* and *ENSCAF00000023289*). The remaining substitution falls in *RNPC3* on chromosome 6.

Table 2 Putatively functional variants fixed for alternative alleles between dogs and wolves

Gene ID	Gene name	Position [chr:position]	Nucleotide change [Dog/Wolf]	Predicted effect
<i>FGF13</i>	Fibroblast growth factor 13	X:108729524	C/G	5'-UTR
<i>FHL1</i>	Four and a half LIM domains 1	X:106604107	A/G	3'- UTR
<i>F9</i>	coagulation factor IX	X:109533147	C/A	3'- UTR
<i>MAP7D3</i>	MAP7 domain containing	X:106609169	C/T	3'- UTR
<i>MBP</i>	Myelin basic protein	1:2951693	G/C	3'- UTR
<i>MCF2</i>	MCF.2 cell line derived transforming sequence	X:109544224	G/C	3'- UTR
<i>RELT</i>	Relt tumor necrosis factor receptor	21:24836981	G/A	3'- UTR
<i>RNPC3</i>	RNA-binding region containing 3	6:47026666	T/G	Missense [T/P]
<i>RNPC3</i>	RNA-binding region containing 3	6:47035497	A/C	Splice region, intronic
<i>SLC9A6</i>	Solute carrier family 9, subfamily A	X:106463600	T/C	3'- UTR
Novel protein coding	ENSCAFG00000018988	X:108560105	T/C	Missense [I/T]
Novel protein coding	ENSCAFG00000018988	X:108560351	A/G	Missense [Q/R]
Novel protein coding	ENSCAFG00000018988	X:108560422	G/A	Missense [E/K]
Novel protein coding	ENSCAFG00000018988	X:108560629	A/G	Missense [M/V]
Novel protein coding	ENSCAFG00000023289	X:77456592	G/A	Missense [E/K]

Fixed variants potentially contributing to behavioral differences

Three of the 11 genes with putatively functional variants fixed for alternative alleles between dogs and wolves are involved in brain development and may therefore potentially contribute to the behavioral differences between dogs and wolves. Of the six genes in the 1Mb candidate sweep region we detect on chromosome one only one gene has a putatively functional variant fixed between dogs and wolves. The gene, *MBP*, encodes myelin basic protein and the segregating site occurs in the 3'-UTR. Myelin basic protein is a component of the myelin sheath, which influences the velocity of axonal impulse conduction [43]. Socially isolated mice show deficits of myelination in the prefrontal cortex, suggesting that myelination is sensitive to behavioral changes [39]. Furthermore, children with autism are significantly more likely to produce anti-MBP antibodies than controls [44].

Intriguingly, another gene that is highly expressed in myelinated nerve fibers [45], *FGF13*, is fixed between dogs and wolves for a putatively functional segregating site in its 5'-UTR. *FGF13* encodes fibroblast growth factor 13 and is within the 500kb region with the second strongest signal of population divergence between dogs and wolves (Table 1). *FGF13* is a growth factor involved in neuronal migration in the cerebral cortex during development [46]. Overexpression of *FGF13* in neuronal cultures from rat embryonic cortex increases the number of neurons containing gamma-aminobutyric acid (GABA) [47], which is notable for the important role of GABA in the regulation of behaviour, including fear [47] and aggression [48]. The presence of a fibroblast growth factor in our list of candidates is potentially supportive of the 'domestication syndrome' hypothesis, which predicts that many of the traits observed in domestic animals are the result of selection on genes related to embryonic development, including fibroblast growth factors [49]. Which of these phenotypes, if any, were targeted by selection will require further investigation.

Perhaps the most intriguing variant fixed between dogs and wolves occurs in the 3'-UTR of *SLC9A6*, which encodes sodium/hydrogen exchanger protein 6. This protein regulates the endoluminal pH of early and recycling endosomes involved in the trafficking of proteins essential for the plasticity of glutamatergic neurons [50]. Loss of function mutations in this gene in humans can lead to Christianson syndrome, also known as "Angelman-like syndrome" [51]. Phenotypes typical of patients with loss of function mutations in *SLC9A6* include cognitive developmental delays, absence of speech, stereotyped repetitive hand movements, hyperkinetic movements and postnatal microcephaly with a narrow

face [51, 52]. Christianson syndrome is also frequently characterised by a happy disposition with easily provoked laughter and smiling, an open mouth with excessive drooling and frequent visual fixation on hands [51, 52]. Several of these phenotypes resemble those that distinguish dogs from wolves. Therefore it is tempting to speculate that selection on regulatory variation influencing expression of *SLC9A6* may have played an important role in producing some of the behavioral phenotypes that emerged during dog domestication.

Variants potentially contributing to anatomical differences

Dogs and wolves are also anatomically distinct [53]. One gene we detect with a variant in the 3'-UTR fixed for alternative alleles between dogs and wolves is *FHL1*, which encodes Four and a half LIMB domains 1. *FHL1* is most highly expressed in skeletal muscle [54]. Defects in this gene in humans result in a variety of muscle disorders, for example scapulo-peroneal myopathy, characterized by progressive weakening of shoulder and lower leg muscles [55, 56]. Selection on this gene may have contributed to the reduced efficiency of skeletal musculature that has been observed in dogs relative to wolves.

Another gene potentially contributing to morphological differences between dogs and wolves is *RNPC3*, which encodes the protein RNA-binding region containing 3. *RNPC3* is involved in pre-mRNA U12-dependent splicing. *RNPC3* is one of only two genes with more than one putatively causal variants fixed between dogs and wolves, the other is a gene of unknown function (Table 2). One variant causes a non-synonymous change while the other is in a predicted intronic splice site. Notably, *RNPC3* is the only autosomal gene with a non-synonymous substitution segregating between all wolves and dogs. Mutations in this gene in humans cause pituitary related growth hormone deficiencies, potentially by disruption of the growth hormone pathway [57]. This pathway also involves the genes *IGF1* and *IGF1R1*, both are associated with haplotypes influencing body size between dog breeds [16, 58], suggesting that this pathway may have been repeatedly targeted by selection for body size during dog domestication.

Interestingly, *RNPC3* is situated less than 1Mb from *AMY2B*, which it has been argued has been selected for increased copy number in dogs as an adaptation to a starch-rich diet [32]. The close proximity of these two genes suggests that the putatively functional variants in *RNPC3* may have risen as a result of hitchhiking, due to selection on the neighbouring *AMY2B*, or vice versa. It is an intriguing possibility that selection in dogs on *AMY2B* for dietary adaptations could have led to morphological changes through the hitchhiking of non-selected functional alleles in the neighbouring *RNPC3*.

Further work will be necessary to untangle the original targets of selection in this case.

Pathway enrichment suggests selection on behaviour

It is not necessarily the case that fixed phenotypic differences between populations must have a fixed genetic basis, particularly in the case of complex polygenic traits. Therefore, we also looked for variants that are not fixed between dogs and wolves. To do this we identified all single nucleotide positions that were highly differentiated between dogs and wolves ($F_{st} \geq 0.75$). From this list of 199,821 sites we used VEP to identify those which had putatively functional consequences. We identify 848 genes with putatively functional variants showing an allele frequency difference of $\geq 75\%$ between dogs and wolves.

We performed a gene ontology enrichment analysis on these 848 genes using the gene ontology and analysis software PANTHER [59, 60]. The only pathway to show a significant enrichment is the ‘adrenaline and noradrenaline biosynthesis pathway’ (P -value = $4.19E-08$) (Table 3). Given the key role of adrenaline in the fight-or-flight response [61] and noradrenaline’s key role as a hormone and neurotransmitter responsible for vigilant attention [62] it is possible that this is driven by genes that have been targeted by selection for changes in behaviour, such as tameness, during dog domestication.

The enrichment signal is the result of putatively functional variants in nine genes (Table 4), including the monoamine oxidases *MAOA* and *MAOB*. The proteins encoded by these genes are involved in the deamination of dopamine, serotonin, adrenaline and noradrenaline. In humans variants in *MAOA* have been associated with aggression [63]. Inhibition of *MAOA* and *MAOB* during brain development induces pathological aggressive behaviour in mice [64] and transgenic mice deficient for *MAOA* show aggressive behaviour and alterations in levels of noradrenaline in the brain [65]. Another gene we identify

is *TH*, which encodes tyrosine hydroxylase, the rate-limiting enzyme in the synthesis of dopamine and noradrenaline [66]. Tyrosine hydroxylase catalyzes the conversion of L-Tyrosine into L-Dopa. Startlingly, the gene encoding DOPA decarboxylase (Aromatic-L-Amino-Acid decarboxylase), which transforms L-Dopa into dopamine, also has a putatively functional variant segregating at high frequency between dogs and wolves (Table 4). This gene, *DDC*, is also involved in several other decarboxylation reactions related to neurotransmitter synthesis, including the conversion of 5-HTP to serotonin [67]. Both *DDC* and *MAOB* have been associated with attention-deficit/hyperactivity disorder in humans [68]. We also detect putative functional variants segregating at high frequency in three genes which encode neurotransmitter transporters in the solute carrier 6 family (*SLC6*). Proteins in the *SLC6* family are involved in the plasma membrane transport of dopamine, noradrenaline, serotonin and GABA and are involved in neurotransmitter signaling [69]. Overall these results strongly suggest that there has been selection for changes in neurotransmitter metabolism during dog domestication, particularly in the catecholamine biosynthesis and transport pathways, which include dopamine, adrenaline and noradrenaline.

Strikingly, polymorphisms in three of these genes have previously been associated with aggressive behaviour within (*SLC6A3* [24]) or between (*TH* [25], *MAOB* [22]) dog breeds. However the alleles in these studies differ from those that we identify. This suggests that the catecholamine pathway has been recurrently targeted by selection during the process of dog domestication. Furthermore, some genes in this pathway show evidence of being recurrently selected during the process of dog domestication, with some variants contributing to behavioral differences between dogs and wolves and others to differences between dog breeds.

Table 3 Panther pathways gene enrichment analysis of genes containing variants with an F_{st} score ≥ 0.75

PANTHER Pathways	Canis familiaris - REFLIST [19662]	Genes with putatively functional variants [848]	Expected number	Fold enrichment	P -value
Unclassified	17318	712	746.91	-	0.00E00
Adrenaline and noradrenaline biosynthesis	26	9	1.12	+	4.19E-08
Axon guidance mediated by netrin	38	7	1.64	+	2.31E-01
Dopamine receptor mediated signaling pathway	51	8	2.20	+	2.97E-01
Nicotine pharmacodynamics pathway	31	6	1.34	+	3.87E-01
Alpha adrenergic receptor signaling pathway	22	5	.95	+	4.45E-01
Gonadotropin releasing hormone receptor pathway	225	19	9.70	+	7.74E-01

Table 4 Putatively functional variants with an F_{st} score ≥ 0.75 in genes in the 'adrenaline and noradrenaline biosynthesis' pathway

Gene ID	Gene name	Position [chr:position]	Nucleotide change [Dog/Wolf]	Predicted effect
<i>DDC</i>	Dopa decarboxylase [Aromatic-L-Amino-Acid decarboxylase]	18:1806717	C/T	Missense [R/Q]
<i>MAOA</i>	Amine oxidase [flavin-containing] A	X:37747023	T/C	3'-UTR
<i>MAOB</i>	Amine oxidase [flavin-containing] B	X:37766049	C/T	3'-UTR
<i>SLC6A3</i>	Sodium-dependent dopamine transporter member 3	34:11239621	C/T	Splice region, intronic
<i>SLC6A17</i>	Sodium-dependent neutral amino acid transporter member 17	6:41776709	C/A	3'-UTR
<i>SLC6A19</i>	Sodium-dependent neutral amino acid transporter member 19	34:11329939	G/A	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30614788	C/T	Missense [S/N]
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30607948	G/A	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30607975	T/C	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30608088	T/C	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30608209	G/A	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30608212	T/C	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30608354	A/G	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30608375	C/T	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30608864	A/G	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:30608989	C/T	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:37750228	T/A	3'-UTR
<i>SNAP29</i>	Synaptosomal-associated protein 29	26:38554416	G/A	3'-UTR
<i>STX7</i>	Syntaxin-7	1:25559797	T/C	3'-UTR
<i>TH</i>	Tyrosine 3-monooxygenase	18:46331581	G/A	Splice region, intronic

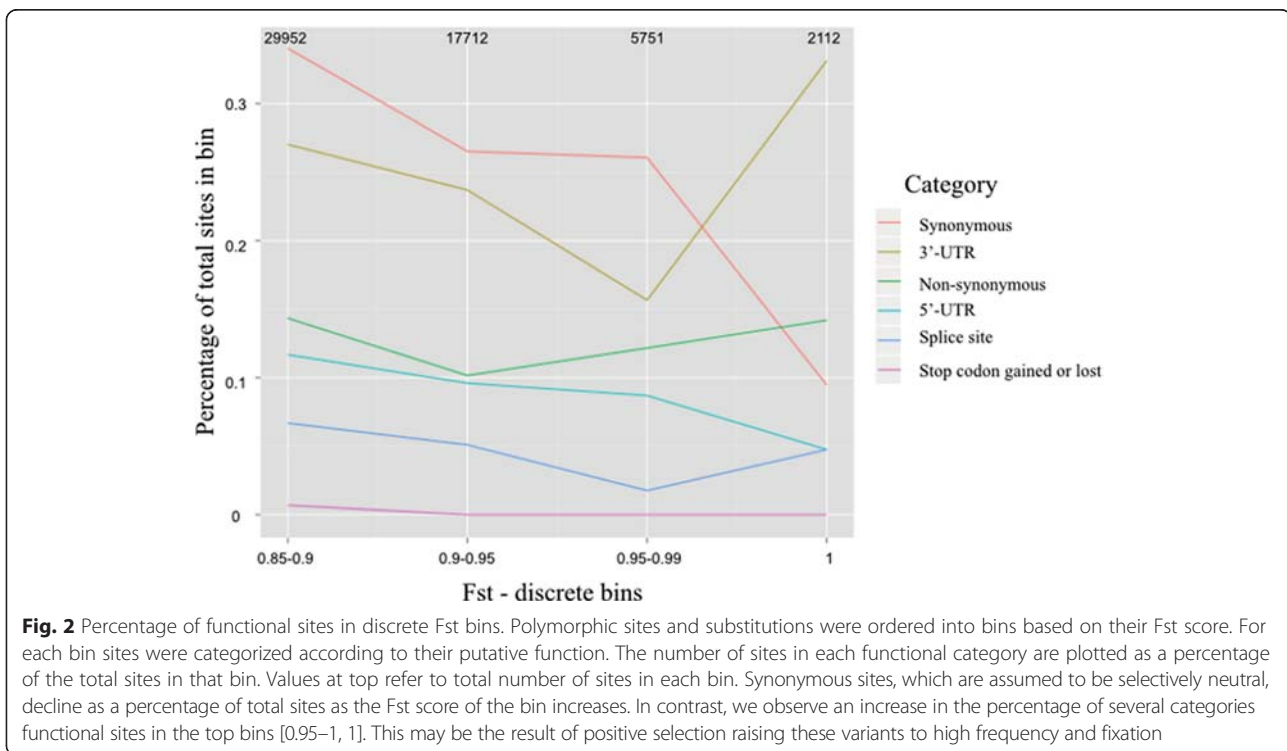
We note that a previous study by Li et al. [35] identified genes involved in glutamate metabolism as the most highly diverged between dogs and wolves. We do not detect this signal in our analysis. This may be partially due to the larger sample size in our study (78 compared to 13 canid genomes), which gives us greater power to detect variants that are truly highly diverged between dogs and wolves. Another explanation is that the analysis of Li et al. [35] was designed to identify genes with highly divergent SNPs irrespective of whether they contain putatively functional variants. Therefore, there may indeed be selection on glutamate metabolism genes in dogs, but the selected variants may reside in nearby regulatory elements. This is supported by their finding that there are gene expression changes in these genes between dogs and wolves [35].

In contrast, our analysis was designed to identify genes with highly divergent putatively functional variants within, or neighbouring, exonic sequences. Therefore, the differing results could be due to selection on the 'adrenaline and noradrenaline biosynthesis pathway' occurring via modifications to the protein structure (missense mutations in *DDC* and *SNAP29*) and flanking proximal regulatory regions (5'-UTR, 3'-UTR and intronic splice sites) of selected genes. While selection on glutamate metabolism may have primarily occurred via

selection on more distal regulatory elements, such as enhancers, potentially influencing tissue specific gene expression. Given the highly polygenic nature of domestication [70], it is plausible that both these pathways have been targeted by selection during dog domestication.

Characterizing the frequency distribution of putatively selected variants

It has been proposed that animal domestication is highly polygenic and can be achieved by the concordant increase in allele frequency of multiple variants without fixation at any loci [70]. We ordered putatively selected sites into bins based on their F_{st} score [0.85–0.9, 0.9–0.95, 0.95–1]. For each discrete bin sites were further categorized based on their putative functional consequences using VEP. The percentage of sites in each functional category are plotted for each bin as a percentage of total sites in that bin (Fig. 2). In the absence of positive selection we expect the proportion of putatively functional variants to decrease as F_{st} increases because purifying selection should act to prevent deleterious mutations rising in frequency [71]. Indeed, for F_{st} values between 0.85–0.95 we see the proportion of all categories of putatively functional sites decreasing as F_{st} increases (Fig. 2). However, for F_{st} values >0.95 we see an increase



in the percentage of several categories of putatively functional sites, particularly sites in the 3'-UTR of genes, while the percentage of synonymous sites, which are presumed to be selectively neutral, decreases. This is suggestive of positive selection acting to bring these variants to fixation.

Evidence that the strength of selection varies around different categories of sites

To further investigate whether selection has preferentially acted on any specific functional categories of sites we calculated mean Fst in 50kb windows centered on each putatively functional variant with an Fst score ≥ 0.75 . Figure 3 shows the distribution of mean Fst around the difference categories of sites, with synonymous variants acting as a control as we do not expect positive selection to be acting on synonymous sites, although this assumption may not always be valid [72]. An ANOVA reveals a significant effect of functional category on mean Fst around sites, $F[5, 2818] = 10.98$, $p = 1.71e-10$ (Additional file 2: Table S1). To find which categories are significantly different we performed Tukey's range test. Although mean Fst is highest around sites that cause a gain of stop codon this is not significantly different as there are only three such sites. We find that non-synonymous variants are in regions of significantly elevated Fst compared to synonymous variants, an observation consistent with positive selection acting on non-synonymous sites (Additional file 3: Table S2).

Interestingly, both synonymous and non-synonymous variants appear to be in regions of significantly higher Fst than variants in the 3'-UTRs and 5'-UTRs. This suggests that during dog domestication selection may have been strongest around non-synonymous variants. However, there are more non-coding than coding variants segregating at high frequency between dogs and wolves, so the overall contribution of each type of variant may still be similar. The elevated mean Fst around synonymous sites relative to regulatory variants may be the result of hitchhiking of synonymous sites that are on the same haplotype as selected variants or, less plausibly, selection on synonymous sites.

Conclusions

Using genome-wide polymorphism data from dogs and wolves we were able to identify putatively functional variants that may have been selected during dog domestication. While previous genomic studies of dog domestication have identified putatively selected regions and genes, this is the first study to combine scans for selection with a genome-wide analysis of multiple categories of putatively functional variants in order to identify specific genetic changes underlying the phenotypic differences between dogs and wolves. We find there are only 11 genes with putatively functional substitutions differentiating all dogs and wolves. Although we note this is likely to be an under-estimate due to our currently limited ability to identify functional variation in non-genic

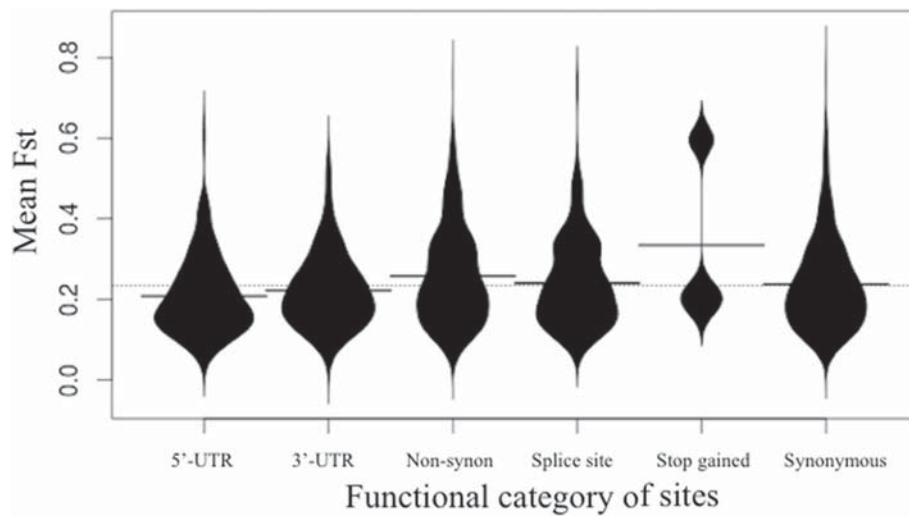


Fig. 3 Distribution of mean F_{st} in 50kb windows centered around putatively functional sites. Polymorphic sites with $F_{st} \geq 0.75$ between dogs and wolves were classified according to their putative function. For each putatively functional site the mean F_{st} was calculated in a 50kb window centered on the site. A violin plot shows the distribution of mean F_{st} values for each category of functional site (5'-UTR, 3'-UTR, non-synonymous, splice site, stop gained, synonymous). Synonymous mutations were included as a category to show the expectation in the absence of positive selection

regions of the genome. The 11 genes that we detect with fixed functional differences between dogs and wolves point towards selection on both morphological and behavioral phenotypes.

We find that, although the majority of putatively functional variants segregating between dogs and wolves are in regulatory regions, in general variants influencing protein structure show the strongest signatures of selection. Although we note that our analysis was restricted to regulatory regions in close proximity to genes. In the future, characterizing the functional effects of these variants may help to further our understanding of the domestication process.

The majority of variants that we detect segregating between dogs and wolves are not fixed but may nevertheless contribute to differences between dogs and wolves due to the polygenic nature of most phenotypes. We provide the first evidence for polygenic selection on putatively functional variation in genes in the adrenaline and noradrenaline biosynthesis pathway during dog domestication. The genes we find implicated in this pathway are involved in the synthesis, transport and degradation of a variety of neurotransmitters, particularly the catecholamines, which include dopamine and noradrenaline. The strong signal of recurrent selection on this pathway and its role in emotional processing and the fight-or-flight response suggests that the behavioral changes we see in dogs compared to wolves may in part be due to changes in this pathway. Furthermore, several of the genes contributing to the signal of enrichment in this pathway have also been associated with levels of

aggressive behaviour between dog breeds [22, 25], suggesting that some of these genes have been important during both the initial domestication process and later breed formation. We note that although the high allele frequency differences between dogs and wolves suggest that the variants we identify were involved in the early domestication process, it is possible that the allelic differentiation we observe occurred later. Looking ahead, ancient DNA from dogs and wolves may provide the temporal resolution to determine which alleles were involved in the earliest stages of dog domestication.

Methods

Data & samples

We used the DoGSD, a publicly available database which contains whole-genome SNP data from dogs and wolves conglomerated from several different studies [37]. All data were obtained from this database and no animal experiments were conducted. For comparability between datasets DoGSD applies a unified variant calling pipeline to all the samples. Using this dataset we analyzed whole-genome variant data from 67 dog and 7 wolf samples (Additional file 4: Table S3), which we treated as two separate groups. The strong genetic drift caused by breed specific population bottlenecks associated with breed creation has resulted in the random fixation of large genomic regions [73]. These could be misidentified as signals of selection. However, we are interested in variants that were selected for during the early domestication process, before the creation of modern breeds. By combining data from as many dogs as possible, from

both modern breeds and village dog populations, we hope to alleviate this problem. Basing our analysis on the reasonable assumption that dog domestication had a single origin [1], we expect variants that were strongly selected for during the early domestication process to be shared across dog breeds, regardless of their more recent population history. While the neutral regions that underwent fixation during breed formation are not expected to be shared across all breeds due to the random nature of genetic drift. Although we note that some variants that were selected for during the early domestication process could be absent from some breeds due to drift from strong bottlenecks associated with the breed creation process.

We excluded the dingo (*Canis lupus dingo*) because although they are now wild, they are thought to be descended from a domesticated Asian dog population [74], which could lead to false negative results if they still contain alleles that were selected for during the early domestication process. To visualize the relationship between samples we created a PCA plot of the samples included in all analyses using EIGENSOFT and SMARTPCA [75, 76] (Additional file 5: Figure S1). The first principal component in the PCA plot clearly differentiates wolves and domestic dogs into two groups. The second principal component appears to differentiate dogs based on their Asian and European ancestry. To reduce the potential for false positives due to low power we only considered sites with genotype calls for $\geq 50\%$ of samples among both the dogs and the wolves.

Genomic scan for selection

To identify regions of the genome with putative signatures of positive selection in dogs or wolves we calculated mean F_{st} across the genome between dogs and wolves in non-overlapping 500kb windows using VCFtools [77]. This is an implementation of Weir and Cockerham's F_{st} [78]. Under neutrality we expect the distribution of mean F_{st} scores to follow a normal distribution. However a histogram of mean F_{st} scores shows a long tail towards positive F_{st} scores, potentially indicative of positive selection (Additional file 6: Figure S2).

Pathway enrichment analysis

Pathway enrichment analysis was performed using the gene ontology and analysis software PANTHER [59–60]. We performed the statistical overrepresentation test using the *Canis familiaris* background gene set and applied the bonferroni correction for multiple hypothesis testing.

Identification of putatively functional sites

The majority of genomic variants are expected to have no impact on the phenotype of an organism. To identify

the putatively functional sites that may have been targeted by selection we used Ensembl's Variant Effect Predictor [VEP] [41]. The VEP predicts the effect of genomic variants on genes, protein sequence and regulatory regions. We classify as putatively functional any sites that influence protein structure; cause missense mutations, frameshifts, or gain or loss of stop codons, and variants that may influence gene expression by being within a 5'-UTR, 3'-UTR, or predicted splice site. While this categorization is likely to be overly conservative, by excluding potentially regulatory variants not situated in or near genes, it will reduce the number of false positives by only including variants with a high probability of having functional consequences.

Coalescent simulations

To test whether the putatively selected 500kb windows with elevated mean F_{st} between dogs and wolves could be the result of a selectively neutral demographic history we performed coalescent simulations with the software *scrm* [79]. The parameters for the simulations were taken from the papers where the samples were first presented. Specifically, we adapted the demographic model presented in [1] (Supplementary Text 8, Command Line 1 *G-PhoCS* model with the full set of migration bands inferred) and incorporated demographic information from the papers where the additional samples were presented [34, 80]. We simulated 148 500kb haplotypes 6000 times, to provide a distribution of regions approximating the dog genome in size. The exact command line is presented in Additional file 7: Table S4. For each simulation we calculated the mean F_{st} of the 500kb haplotypes between dogs and wolves using the R package PopGenome [81].

Availability of supporting data

The dataset supporting the conclusions of this article is available in the DoGSD repository [37] [<http://dogsd.big.ac.cn/snp/pages/download/download.jsp>].

Additional files

Additional file 1: Figure S3. Mean F_{st} of 500kb regions. Distribution of the empirical data compared to results obtained from coalescent simulations. The empirical distribution is presented both with (red line) and without the regions from the X chromosome (blue line). The long tail of the empirical data is absent in the neutral simulations, suggesting that positive selection may explain the elevated F_{st} in these regions. (DOCX 73 kb)

Additional file 2: Table S1. Results of ANOVA of mean F_{st} in 50kb windows around functional categories of sites with $F_{st} \geq 0.75$. (DOCX 41 kb)

Additional file 3: Table S2. Results of Tukey's range test for ANOVA of Mean F_{st} in 50kb windows around functional categories of sites with $F_{st} \geq 0.75$. (DOCX 99 kb)

Additional file 4: Table S3. Samples from the DoGSD included in this study. (DOCX 121 kb)

Additional file 5: Figure S1. PCA plot of samples included in this study. PCA of genome-wide polymorphism data from 67 dogs and 7 wolves. The percentage of the total variance explained by the first and second principal component are labeled on the X and Y axis, respectively. PC1 clearly separates dogs from wolves while PC2 primarily separates dogs by geographic origin. (DOCX 144 kb)

Additional file 6: Figure S2. Histogram of mean Fst scores calculated in 500kb windows genome-wide between dogs and wolves. Histogram of mean Fst calculated in 500kb genomic windows across the autosome and X chromosome between dogs and wolves. Counts are included above each bin. The long tail towards positive mean Fst scores is potentially indicative of positive selection. (DOCX 71 kb)

Additional file 7: Table S4. The scrm command line used for coalescent simulations of dog and wolf demographic history and Ne estimates and parameters used for the simulations. (DOCX 77 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AC computed the analyses with contributions from TB. AC conceived the study and wrote the manuscript. Both authors participated in reading and approving the final manuscript.

Acknowledgments

We thank the Max Planck Society for making this research possible. We thank S. Pääbo for constructive criticism of the manuscript. We thank G. Wang for providing early access to the data. We thank S. Peyrégne for help with coalescent simulations. We also thank anonymous reviewers for their helpful comments.

Received: 11 August 2015 Accepted: 22 December 2015

Published online: 12 January 2016

References

- Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, et al. Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genet.* 2014;10:e1004016.
- Davis SJM, Valla FR. Evidence for domestication of the dog 12,000 years ago in the Natufian of Israel. *Nature.* 1978;276:608–10.
- Wayne RK, Ostrander EA. Lessons learned from the dog genome. *Trends Genet.* 2007;11(23):557–67.
- Hare B, Tomasello M. Human-like social skills in dogs? *Trends Cogn Sci.* 2005;9:439–44.
- Hare B, Plyusnina I, Ignacio N, Schepina O, Stepika A, Wrangham R, et al. Social cognitive evolution in captive foxes is a correlated by-product of experimental domestication. *Curr Biol.* 2005;15:226–30.
- Range F, Virányi Z. Tracking the evolutionary origins of dog-human cooperation: the "Canine Cooperation Hypothesis". *Front Psychol.* 2015;5:1582.
- Serpell J, Duffy D. Dog Breeds and Their Behavior. In: *Domestic Dog Cognition and Behavior.* Berlin, Heidelberg: Springer; 2014.
- Spady TC, Ostrander EA. Canine behavioral genetics: pointing out the phenotypes and herding up the genes. *Am J Hum Genet.* 2008;1(82):10–8.
- Stockard C, James W. The genetic and endocrinic basis for differences in form and behavior: as elucidated by studies of contrasted pure-line dog breeds and their hybrids. Philadelphia: The Wistar Institute of Anatomy and Biology; 1941.
- Scott JP, Fuller JL: Genetics and the Social Behavior of the Dog. Cambridge: University of Chicago Press; 1965:468
- Lark KG, Chase K, Sutter NB. Genetic architecture of the dog: sexual size dimorphism and functional morphology. *Trends Genet.* 2006;22:537–44.
- Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, et al. A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol.* 2010;8:49–50.
- Pollinger JP, Bustamante CD, Fedel-Alon A, Schmutz S, Gray MM, Wayne RK. Selective sweep mapping of genes with large phenotypic effects. *Genome Res.* 2005;15:1809–19.
- Haworth KE, Islam I, Breen M, Putt W, Makrinou E, Binns M, et al. Canine TCOF1; cloning, chromosome assignment and genetic analysis in dogs with different head types. *Mamm Genome.* 2001;12:622–9.
- Mosher DS, Quignon P, Bustamante CD, Sutter NB, Mellersh CS, Parker HG, et al. A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs. *PLoS Genet.* 2007;3:779–86.
- Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, Zhu L, et al. A single IGF1 allele is a major determinant of small size in dogs. *Sci [New York, NY].* 2007;316:112–5.
- Rimbault M, Ostrander EA. So many doggone traits: Mapping genetics of multiple phenotypes in the domestic dog. *Hum Mol Genet.* 2012;21:R52–7.
- Houpt KA. Genetics of canine behavior. *Acta Vet Brno.* 2007;3(76):431–44.
- Rigterink A, Houpt K, Cho M, Eze O. Genetics of canine behavior: A review. *World J Med Genet.* 2014;4(3):46–57.
- Trut L. Early canid domestication: the farm-fox experiment. *Am Sci.* 1999;2(87):160–69.
- Copping R, Schneider R: Evolution of working dogs. The domestic dog: Its evolution, behaviour and interactions with people. Cambridge: Cambridge University press, 1995.
- Hashizume C, Masuda K, Momozawa Y, Kikusui T, Takeuchi Y, Mori Y. Identification of a cysteine-to-arginine substitution caused by a single nucleotide polymorphism in the canine monoamine oxidase B gene. *J Vet Med Sci.* 2005;67:199–201.
- Ito H, Nara H, Inoue-Murayama M, Shimada MK, Koshimura A, Ueda Y, et al. Ito Shin'ichi: Allele Frequency Distribution of the Canine Dopamine Receptor D4 Gene Exon III and I in 23 Breeds. *J Vet Med Sci.* 2004;66:815–20.
- Lit L, Belanger JM, Boehm D, Lybarger N, Haverbeke A, Diederich C, et al. Characterization of a dopamine transporter polymorphism and behavior in Belgian Malinois. *BMC Genet.* 2013;14:45.
- Takeuchi Y, Hashizume C, Chon EMH, Momozawa Y, Masuda K, Kikusui T, et al. Canine tyrosine hydroxylase [TH] gene and dopamine beta-hydroxylase [DBH] gene: their sequences, genetic polymorphisms, and diversities among five different dog breeds. *J Vet Med Sci.* 2005;67:861–7.
- Våge J, Wade C, Biagi T, Fatjó J, Amat M, Lindblad-Toh K, et al. Association of dopamine- and serotonin-related genes with canine aggression. *Genes Brain Behav.* 2010;9:372–8.
- Saetre P, Lindberg J, Leonard JA, Olsson K, Pettersson U, Ellegren H, et al. From wild wolf to domestic dog: gene expression changes in the brain. *Brain Res Mol Brain Res.* 2004;126:198–206.
- Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, et al. A comparison of brain gene expression levels in domesticated and wild animals. *PLoS Genet.* 2012;8:e1002962.
- Våge J, Bønsdorff TB, Arnet E, Tverdal A, Lingaas F. Differential gene expression in brain tissues of aggressive and non-aggressive dogs. *BMC Vet Res.* 2010;6:34.
- Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, et al. Tracking footprints of artificial selection in the dog genome. *Proc Natl Acad Sci U S A.* 2010;107:1160–5.
- Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, et al. Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. *PLoS Genet.* 2011;7:e1002316.
- Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature.* 2013;495:360–4.
- Li Y, Von Holdt BM, Reynolds A, Boyko AR, Wayne RK, Wu DD, et al. Artificial selection on brain-expressed genes during the domestication of dog. *Mol Biol Evol.* 2013;30:1867–76.
- Wang GD, Zhai W, Yang HC, Fan RX, Cao X, Zhong L, et al. The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat Commun.* 2013;4:1860.
- Li Y, Wang GD, Wang MS, Irwin DM, Wu DD, Zhang YP. Domestication of the dog from the wolf was promoted by enhanced excitatory synaptic plasticity: a hypothesis. *Genome Biol Evol.* 2014;6:3115–21.
- Poh Y-P, Domingues VS, Hoekstra HE, Jensen JD. On the prospect of identifying adaptive loci in recently bottlenecked populations. *PLoS One.* 2014;9:e110579.
- Bai B, Zhao W-M, Tang B-X, Wang Y-Q, Wang L, Zhang Z, et al. DoGSD: the dog and wolf genome SNP database. *Nucleic Acids Res.* 2015;43(Database issue):D777–83.

38. Vicoso B, Charlesworth B. Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet.* 2006;7:645–53.
39. Liu J, Dietz K, DeLoyht JM, Pedre X, Kelkar D, Kaur J, et al. Impaired adult myelination in the prefrontal cortex of socially isolated mice. *Nat Neurosci.* 2012;15:1621–3.
40. Pritchard JK, Di Rienzo A. Adaptation - not by sweeps alone. *Nat Rev Genet.* 2010;11:665–7.
41. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics.* 2010;26:2069–70.
42. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. *PLoS Genet.* 2015;11:e1005004.
43. Sakamoto Y, Kitamura K, Yoshimura K, Nishijima T, Uyemura K. Complete amino acid sequence of PO protein in bovine peripheral nerve myelin. *J Biol Chem.* 1987;262:4208–14.
44. Singh VK, Warren RP, Odell JD, Warren WL, Cole P. Antibodies to myelin basic protein in children with autistic behavior. *Brain Behav Immun.* 1993;7:97–103.
45. Zhang X, Bao L, Yang L, Wu Q, Li S. Roles of intracellular fibroblast growth factors in neural development and functions. *Sci China Life Sci.* 2012;55:1038–44.
46. Greene JM, Li YL, Yourey PA, Gruber J, Carter KC, Shell BK, et al. Identification and characterization of a novel member of the fibroblast growth factor family. *Eur J Neurosci.* 1998;10:1911–25.
47. Makkar SR, Zhang SQ, Cranney J. Behavioral and neural analysis of GABA in the acquisition, consolidation, reconsolidation, and extinction of fear memory. *Neuropsychopharmacology.* 2010;35:1625–52.
48. Almada RC, Coimbra NC. Recruitment of striatonigral disinhibitory and nigroreticular inhibitory GABAergic pathways during the organization of defensive behavior by mice in a dangerous environment with the venomous snake *Bothrops alternatus* [Reptilia , Viperidae]. *Synapse* 2015;n/a–n/a.
49. Wilkins AS, Wrangham RW, Fitch WT. The “Domestication Syndrome” in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics. *Genetics.* 2014;197:795–808.
50. Zanni G, Barresi S, Cohen R, Specchio N, Basel-Vanagaite L, Valente EM, et al. A novel mutation in the endosomal Na⁺/H⁺ exchanger NHE6 [SLC9A6] causes Christianson syndrome with electrical status epilepticus during slow-wave sleep [ESES]. *Epilepsy Res.* 2014;108:811–5.
51. Gilfillan GD, Selmer KK, Roxrud I, Smith R, Kyllerman M, Eiklid K, et al. SLC9A6 mutations cause X-linked mental retardation, microcephaly, epilepsy, and ataxia, a phenotype mimicking Angelman syndrome. *Am J Hum Genet.* 2008;82:1003–10.
52. Schroer RJ, Holden KR, Tarpey PS, Matheus MG, Griesemer DA, Friez MJ, et al. Natural history of Christianson syndrome. *Am J Med Genet A.* 2010;152A:2775–83.
53. Drake AG, Coquerelle M, Colombeau G. 3D morphometric analysis of fossil canid skulls contradicts the suggested domestication of dogs during the late Paleolithic. *Sci Rep.* 2015;5:8299.
54. Lee SMY, Tsui SKW, Chan KK, Garcia-Barcelo M, Wayne MMY, Fung KP, et al. Chromosomal mapping, tissue distribution and cDNA sequence of Four-and-a-half LIM domain protein 1 [FHL1]. *Gene.* 1998;216:163–70.
55. Quinzii CM, Vu TH, Min KC, Tanji K, Barral S, Grewal RP, et al. X-linked dominant scapuloperoneal myopathy is due to a mutation in the gene encoding four-and-a-half-LIM protein 1. *Am J Hum Genet.* 2008;82:208–13.
56. Chen D-H, Raskind WH, Parson WW, Sonnen JA, Vu T, Zheng Y, et al. A novel mutation in FHL1 in a family with X-linked scapuloperoneal myopathy: Phenotypic spectrum and structural study of FHL1 mutations. *J Neurol Sci.* 2010;296:22–9.
57. Argente J, Flores R, Gutiérrez-Arumí A, Verma B, Martos-Moreno GÁ, Cuscó I, et al. Defective minor spliceosome mRNA processing results in isolated familial growth hormone deficiency. *EMBO Mol Med.* 2014;6:299–306.
58. Hoopes BC, Rimbault M, Liebers D, Ostrander EA, Sutter NB. The insulin-like growth factor 1 receptor [IGF1R] contributes to reduced size in dogs. *Mamm Genome.* 2012;23:780–90.
59. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 2003;13:2129–41.
60. Mi H, Thomas P. *Protein Networks and Pathway Analysis.* Volume 563. Totowa, NJ: Humana Press; 2009. p. 123–40 [Methods in Molecular Biology].
61. Engelmamm M, Landgraf R, Wotjak CT. The hypothalamic-neurohypophysial system regulates the hypothalamic-pituitary-adrenal axis under stress: an old concept revisited. *Front Neuroendocrinol.* 2004;25:132–49.
62. Howells FM, Stein DJ, Russell VA. Synergistic tonic and phasic activity of the locus coeruleus norepinephrine [LC-NE] arousal system is required for optimal attentional performance. *Metab Brain Dis.* 2012;27:267–74.
63. Beitchman JH, Mik HM, Ehtesham S, Douglas L, Kennedy JL. MAOA and persistent, pervasive childhood aggression. *Mol Psychiatry.* 2004;9:546–7.
64. Mejia JM, Ervin FR, Baker GB, Palmour RM. Monoamine oxidase inhibition during brain development induces pathological aggressive behavior in mice. *Biol Psychiatry.* 2002;52:811–22.
65. Cases O, Seif I, Grimsby J, Gaspar P, Chen K, Pournin S, et al. Aggressive behavior and altered amounts of brain serotonin and norepinephrine in mice lacking MAOA. *Sci [New York, NY].* 1995;268:1763–6.
66. Nagatsu T, Levitt M, Udenfriend S. Tyrosine Hydroxylase. The initial step in norepinephrine biosynthesis. *J Biol Chem.* 1964;239:2910–7.
67. Lovenberg W, Weissbach H, Udenfriend S. Aromatic LAmho acid decarboxylase. *J Biol Chem.* 1962;1(237):89–93.
68. Ribasés M, Ramos-Quiroga JA, Hervás A, Bosch R, Bielsa A, Gastaminza X, et al. Exploration of 19 serotonergic candidate genes in adults and children with attention-deficit/hyperactivity disorder identifies association for 5HT2A, DDC and MAOB. *Mol Psychiatry.* 2009;14:71–85.
69. Chen N-H, Reith MEA, Quick MW. Synaptic uptake and beyond: the sodium- and chloride-dependent neurotransmitter transporter family SLC6. *Pflügers Arch.* 2004;447:519–31.
70. Carneiro M, Rubin C-J, Di Palma F, Albert FW, Alföldi J, Barrio AM, et al. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Sci [New York, NY].* 2014;345:1074–9.
71. Nielsen R. Molecular signatures of natural selection. *Annu Rev Genet.* 2005;39:197–218.
72. Lawrie DS, Messer PW, Hershberg R, Petrov DA. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet.* 2013;9:e1003527.
73. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature.* 2005;438:803–19.
74. Savolainen P, Leitner T, Wilton AN, Matisoo-Smith E, Lundeberg J. A detailed picture of the origin of the Australian dingo, obtained from the study of mitochondrial DNA. *Proc Natl Acad Sci U S A.* 2004;101:12387–90.
75. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick N, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38:904–9.
76. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006;2:2074–93.
77. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinform.* 2011;27:2156–8.
78. Weir B, Cockerham C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution.* 1984;38:1358–70.
79. Staab PR, Zhu S, Metzler D, Lunter G. scm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics.* 2015;31(10):1680–2. doi:10.1093/bioinformatics/btu861.
80. Gou X, Wang Z, Li N, Qiu F, Xu Z, Yan D, et al. Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome Res.* 2014;24:1308–15.
81. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. PopGenome: An efficient swiss army knife for population genomic analyses in R. *Mol Biol Evol.* 2014;31:1929–36.

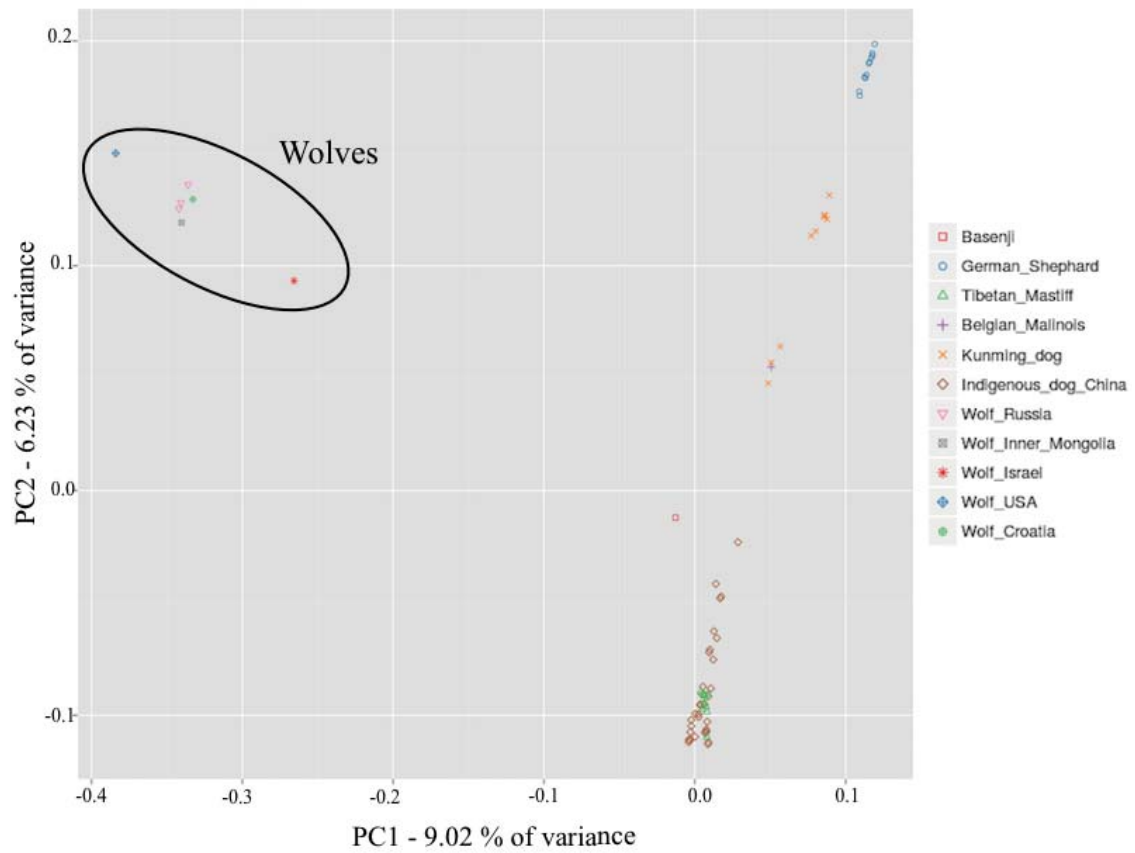
Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

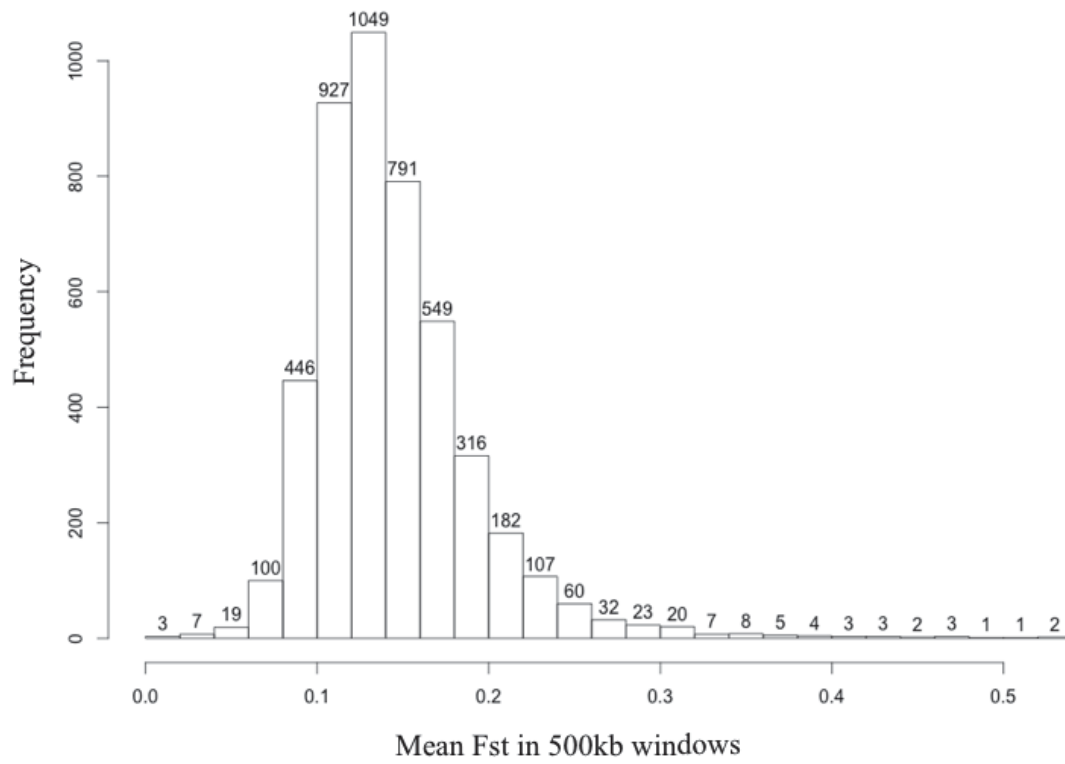
Submit your manuscript at
www.biomedcentral.com/submit



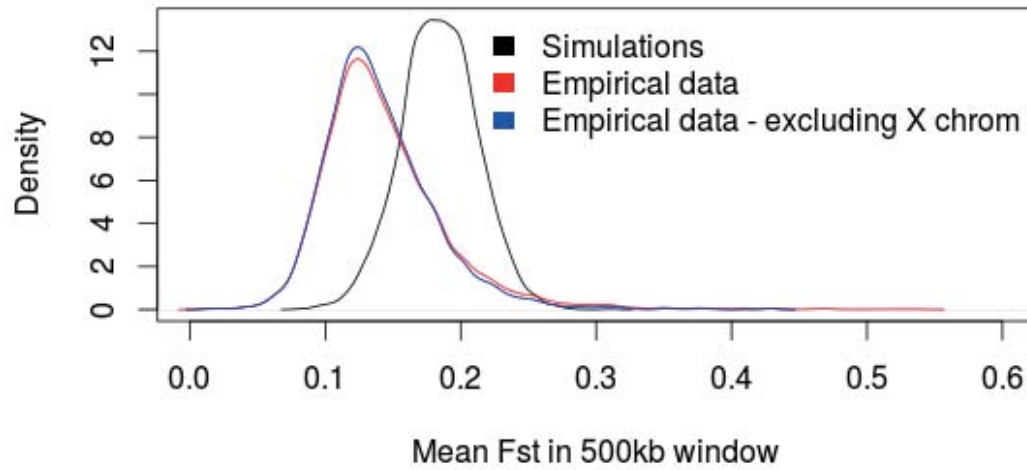
Supplementary Figure 1. PCA of samples.



Supplementary Figure 2. Distribution of 500kb mean Fst scores.



Supplementary Figure 3. Mean Fst of 500kb regions. Distribution of the empirical data compared to results obtained from coalescent simulations.



Supplementary Table 1. Results of ANOVA of mean Fst in 50kb windows around functional categories of sites with Fst \geq 0.75

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Functional categories	5	0.63	0.12544	10.98	1.71e-10
Residuals	2818	32.18	0.01142		

Supplementary Table 2. Results of Tukey's range test for ANOVA of Mean Fst in 50kb windows around functional categories of sites with Fst \geq 0.75

Comparison	Difference in observed means	Lower interval	Upper interval	Adjusted P-value
5'-UTR-3'-UTR	-0.014384125	-0.037221200	0.008452951	0.4683529
Non_synonymous-3'-UTR	0.035246115	0.018772050	0.051720181	0.0000000
Splice_site-3'-UTR	0.018374072	-0.007359839	0.044107983	0.3220012
Stop_gained-3'-UTR	0.110713406	-0.065507770	0.286934581	0.4713171
Synonymous-3'-UTR	0.015509507	0.001511038	0.029507977	0.0198602
Non_synonymous-5'-UTR	0.049630240	0.025256569	0.074003911	0.0000001
Splice_site-5'-UTR	0.032758197	0.001374838	0.064141555	0.0348514
Stop_gained-5'-UTR	0.125097530	-0.052036836	0.302231896	0.3344823
Synonymous-5'-UTR	0.029893632	0.007120004	0.052667260	0.0025500
Splice_site-Non_synonymous	-0.016872043	-0.043978832	0.010234745	0.4822043
Stop_gained-Non_synonymous	0.075467290	-0.100959597	0.251894177	0.8273125
Synonymous-Non_synonymous	-0.019736608	-0.036122607	-0.003350609	0.0079127
Stop_gained-Splice_site	0.092339334	-0.085191750	0.269870418	0.6751021
Synonymous-Splice_site	-0.002864565	-0.028542187	0.022813058	0.9995681
Synonymous-Stop_gained	-0.095203898	-0.271416863	0.081009066	0.6378315

Supplementary Table 3. Samples from the DoGSD included in this study

ID	Breed	Sample Location	Coverage(X)
GS1	German Shepherd	Kunming	15.66
GS2	German Shepherd	Kunming	16.44
GS3	German Shepherd	Kunming	16.07
GS4	German Shepherd	Kunming	15.97
GS5	German Shepherd	Kunming	15.24
GS6	German Shepherd	Kunming	16.27
GS7	German Shepherd	Kunming	16.86
GS8	German Shepherd	Kunming	15.33
GS9	German Shepherd	Kunming	16.93
GS10	German Shepherd	Kunming	15.98
KM1	Indigenous dog	Kunming	12.01
KM2	Indigenous dog	Kunming	17.32
KM3	Indigenous dog	Kunming	16.48
KM4	Indigenous dog	Kunming	16.92
KM5	Indigenous dog	Kunming	16.73
KM6	Indigenous dog	Kunming	16.23
KM7	Indigenous dog	Kunming	18.14
KM8	Indigenous dog	Kunming	14.44
KM9	Indigenous dog	Kunming	14.74
KM10	Indigenous dog	Kunming	17.21
YJ1	Indigenous dog	Yingjiang	16.81
YJ2	Indigenous dog	Yingjiang	15.03
YJ3	Indigenous dog	Yingjiang	14.71
YJ4	Indigenous dog	Yingjiang	14.94
YJ5	Indigenous dog	Yingjiang	14.42
YJ6	Indigenous dog	Yingjiang	15.32
YJ7	Indigenous dog	Yingjiang	15.17
YJ8	Indigenous dog	Yingjiang	15.54
YJ9	Indigenous dog	Yingjiang	15.55
YJ10	Indigenous dog	Yingjiang	16.22
LJ1	Indigenous dog	Lijiang	16.76
LJ2	Indigenous dog	Lijiang	16.72
LJ3	Indigenous dog	Lijiang	15.97
LJ4	Indigenous dog	Lijiang	14.35
LJ5	Indigenous dog	Lijiang	15.28
LJ6	Indigenous dog	Lijiang	15.78
LJ7	Indigenous dog	Lijiang	13.54
LJ8	Indigenous dog	Lijiang	16.03
LJ9	Indigenous dog	Lijiang	16.02
LJ10	Indigenous dog	Lijiang	15.38
DQ1	Indigenous dog	Diqing	16.60
DQ2	Indigenous dog	Diqing	17.23
DQ3	Indigenous dog	Diqing	15.13
DQ4	Indigenous dog	Diqing	15.81
DQ5	Indigenous dog	Diqing	16.48

DQ6	Indigenous dog	Diqing	13.85
DQ7	Indigenous dog	Diqing	17.00
DQ8	Indigenous dog	Diqing	13.27
DQ9	Indigenous dog	Diqing	14.97
DQ10	Indigenous dog	Diqing	16.37
TM1	Tibetan Mastiff	Diqing	16.07
TM2	Tibetan Mastiff	Diqing	14.34
TM3	Tibetan Mastiff	Diqing	14.66
TM4	Tibetan Mastiff	Diqing	13.91
TM5	Tibetan Mastiff	Diqing	14.91
TM6	Tibetan Mastiff	Diqing	15.51
TM7	Tibetan Mastiff	Diqing	16.63
TM8	Tibetan Mastiff	Diqing	14.74
TM9	Tibetan Mastiff	Diqing	16.71
TM10	Tibetan Mastiff	Diqing	16.70
FAMICHN00001	Indigenous dog	Xi'an, China	19.10
FAMICHN00002	Indigenous dog	Simao, China	11.43
FAMICHN00003	Indigenous dog	Ya'an, China	12.63
LUPWRUS00001	Grey wolf	Altai, Russia	11.32
LUPWRUS00002	Grey wolf	Chukotka, Russia	11.59
LUPWRUS00003	Grey wolf	Bryansk, Russia	30.73
LUPWCHN00001	Grey wolf	Inner Mongolia, China	19.05
FAMBGSD00001	German Shepherd Dog	NA	9.53
FAMBTIM00001	Tibetan Mastiff	NA	10.99
FAMBBEM00001	Belgian Malinois	NA	9.97
Basenji	Basenji	Bethesda, MD, USA	4.25
CHW	Wolf	San Diego Zoo, CA, USA	19.62
CRW	Wolf	Perković, Croatia	6.70
ISW	Wolf	Neve Ativ, Golan Heights, Israel	5.04

Supplementary Table 4. The scrm command line used for coalescent simulations of dog and wolf demographic history and Ne estimates and parameters used for the simulations.

```
scrm 148 6000 -t 900 -r 225 500000 -I 9 6 4 2 2 86 22 22 2 2 -n 1 0.12 -n 2 0.12 -n 3
0.253 -n 4 0.581 -n 5 0.378 -n 6 0.0036 -n 7 0.0036 -n 8 0.0036 -n 9 0.0578 -m 4 9
13500 -m 9 4 32400 -m 2 5 2358 -m 1 5 2358 -m 5 2 3078 -m 5 1 3078 -ej 0.0248 2 1
-ej 0.0248 3 1 -ej 0.0248 4 1 -en 0.0248 1 0.2889 -ej 0.000213 8 7 -ej 0.009259 7 6 -ej
0.022407 6 5 -ej 0.022407 9 5 -ej 0.027593 5 1 -em 0.022407 9 4 0 -em 0.022407 4 9
0 -em 0.022407 5 2 0 -em 0.022407 2 5 0 -em 0.022407 1 5 0 -em 0.022407 5 1 0 -en
0.0248 1 0.2889 -en 0.022407 5 0.0433 -en 0.027593 1 1
```

Population	Number of haplotypes	Present day Ne
Russian grey wolf	6	5400
Chinese grey wolf	4	5400
Croatian grey wolf	2	11400
Israeli grey wolf	2	26150
Indigenous dog	86	17000
Tibetan mastiff	22	200
German shephard	22	200
Belgian malinois	2	200
Basenji	2	2600
Mutation rate	1x10E-8 mutations per generation	
Recombination rate	0.25cM/Mb	
Generation time	3 years	

Chapter 2

Genes and pathways selected during animal domestication

Submitted to eLife

by

Alex Cagan, Frank W Albert, Irina Plyusnina, Lyudmila Trut, Gabriel Renaud, Frederic Romagné, Victor Wiebe, Rimma Kozhemjakina, Rimma Gulevich, Oleg Trapezov, Nikolay Yudin, Tatyana Alekhina, Ruslan Aitnazarov, Ludmila Trapezova, Yury Herbeck, Torsten Schöneberg , Svante Pääbo.



Genes and pathways selected during animal domestication

Alex Cagan¹, Frank W. Albert², Irina Plyusnina³, Lyudmila Trut³, Gabriel Renaud¹, Frederic Romagne¹, Victor Wiebe¹, Rimma Kozhemjakina³, Rimma Gulevich³, Oleg Trapezov^{3,4}, Nikolay Yudin^{3,4}, Tatyana Alekhina³, Ruslan Aitnazarov³, Ludmila Trapezova³, Yury Herbeck³, Torsten Schöneberg⁵, and Svante Pääbo¹

¹Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; ²Department of Human Genetics, Cell Biology, & Development, University of Minnesota, Minneapolis, MN; ³The Federal Research Center Institute of Cytology and Genetics, the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia; ⁴Novosibirsk State University, Novosibirsk, Russia; ⁵University of Leipzig, Leipzig, Germany

Abstract

The domestication of a small number of animal species was essential for the emergence of complex human societies. Despite decades of interest, the biological changes underlying the domestication process remain poorly understood. We generated whole-genome sequences from “domesticated” Norway rats and American mink, and identified genes and putatively functional variants that may underlie the phenotypic differences seen in the domesticated animals. When we combine these data with whole-genome sequences from seven pairs of domestic animals and their wild sister species we find six biological pathways that were recurrently affected by the domestication process in all nine domesticated animal species. One of these is the ErbB signaling pathway, involved in the development of the reproductive system and neural crest migration.

INTRODUCTION

The domestication of a few mammalian species had a transformative effect on human societies by providing transport, labor, food, and companionship. Intriguingly, these domesticated species tend to share a suite of behavioral and morphological traits, including tameness, smaller brains, and white coat spotting, sometimes termed 'domestication syndrome' (Hammer, 1984; Sánchez-Villagra et al., 2016). It has been suggested that this may be due to the pleiotropic effects of genetic variants selected for reduced fear and aggression towards humans, or 'tameness', which is likely to be a feature selected early during domestication (Belyaev, 1979; Trut et al., 2009; Wilkins et al., 2014). Several of these traits have also been seen in foxes, rats and mink that have been experimentally selected for tameness towards humans (Belyaev, 1979; Trut et al., 2009). Previous genomic studies found that domestication is a highly polygenic process (Axelsson et al., 2013; Frantz et al., 2015; Carneiro et al., 2014), even if a few genes may have changed expression in several domesticated species (Albert et al., 2012). However, it is unclear whether selection on certain genes and pathways was a common feature in many domesticated animal species or if domestication was fundamentally different in each mammalian species (Larson et al., 2014).

RESULTS

Experimental rat and mink lines

The two experimental lines of Norway rats and American mink have been selected over 70 and 15 generations, respectively, for either a less fearful ("tame") or more fearful ("aggressive") response to humans (Belyaev, 1979; Trut et al., 2009; Naumenko et al., 1989). Selection for tameness produced several unintended behavioral and morphological changes, some of which mirror those seen in domestic animals, including changes in coat-color and skull-shape (Albert et al., 2009; Gulevich et al., 2010; Trut et al., 2004).

We sequenced the 20 tame and 20 aggressive rats and mink, respectively, to ~5-fold genomic depth per individual (Supplementary file 1A, Materials and methods). We aligned the rat sequences to the rat reference genome (*Rattus norvegicus*, rno5) and the mink sequences to the ferret genome (*Mustela putorius furo*, musFur1), the closest reference genome available (Materials and methods). We detect 83,459 nucleotide sites where the tame and aggressive rats are fixed for alternative alleles. In contrast, we find only one fixed difference between the tame and aggressive mink.

To explore differences between these two experiments further we performed principal component analyses (PCA) using the software Eigenstrat and SmartPCA (Patterson et al., 2006; Price et al., 2006), using genome-wide SNPs (Figure 1). We observe that the three rat lines cluster distinctly from one another on the 1st and 2nd principal components (Figure 1A). In contrast, the tame and aggressive mink are separated on the 1st

principal component but not the 2nd, whereas the unselected animals are not full separate from the tame animals (Figure 1B). These differences likely reflect the fact that the mink lines have been selected for only 15 generations, while the rat lines have been selected for 70 generations, resulting in more time for differentiation due to a combination of selection and genetic drift.

Signatures of selection

To identify genes that may have been positively selected in the experimental lines we used an outlier-based approach. Positive selection in one of the populations is expected to increase the local genomic divergence between the lines, as the frequency of selected haplotypes increases. To identify signals of recent selection in the tame and aggressive lines, we therefore calculated the absolute allele frequency difference (ΔAF) of all polymorphic sites and fixed differences (SNPs) in exons, introns and 500bp up- and down-stream of coding regions between the lines. Using these sites we then calculated the mean ΔAF for each gene. To estimate the expected mean ΔAF under neutrality for each gene we used a resampling approach controlling for gene length (Supplementary Figure 1, Materials and methods). This resampling approach assumes that SNPs with different ΔAF s are randomly distributed across the genome, whereas selective sweeps are expected to create a local clustering of high ΔAF SNPs due to hitchhiking. We identify 4,099 and 1,241 genes with a mean ΔAF above the 99th quantile of the distribution obtained from resampling in the rat and mink, respectively (Supplementary files 2A-B). The higher number of diverged genes in the rat presumably reflects the greater number of generations that these lines have been divergently selected.

To gauge if some of these genes may have been selected during domestication we overlap them with quantitative trait loci (QTLs) for tameness and aggression that were previously mapped in an F2 cross of these rat lines (Albert et al., 2009) (Materials and methods). Among the highly diverged genes that are also located in the tameness QTLs, several biological pathways as defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Supplementary file 3A) (Kanehisa *et al.*, 2016; Kanehisa and Goto, 2004) are over-represented (Materials and methods). The most significant of these is the pathway 'Neuroactive ligand-receptor interaction' (p-value = 1.88e-09). Of the 20 genes responsible for this signal 19 encode cell-surface receptors (Supplementary file 4A), many of which are associated with relevant behavioral phenotypes. These include two genes encoding gamma-aminobutyric acid (GABA) receptor A subunits (*Gabrg2*, *Gabra1*). In mice, heterozygous knock-outs of GABA receptor A subunits display hyperactivity of the hypothalamic–pituitary–adrenal (HPA) axis (Shen et al., 2010), an endocrine system involved in the so-called “fight-or-flight response”. Notably, *Gabra1* homozygous knock-out mice have heightened levels of anxiety in behavioral assays, including a stronger auditory startle response (Ye et al., 2010). Another of the 20 genes encodes corticotropin releasing hormone receptor 2 (*Crhr2*), which is expressed in the anterior lobe of the pituitary gland and stimulates the release of adrenocorticotrophic hormone (ACTH). Mice with homozygous or heterozygous deletions of *Crhr2* exhibit enhanced levels of anxiety in several behavioral tests (Kishimoto

et al., 2000). Selection on these gene may contribute to the changes in the responsiveness of the HPA system to restriction stress that have been observed in the tame rats (Shikhevich et al., 2003).

Comparative analysis of selection signatures

To identify genes and pathways that may often - or even always - be involved in domestication, we looked for genes that have diverged drastically not only between the tame and aggressive rats and mink, but also in animals domesticated during human history. To do this, we used whole-genome sequences from populations of seven domesticated animals (Asian pigs, European pigs, cats, dogs, goats, sheep and rabbits) and their closest extant non-domesticated relatives for which whole-genome sequence data are available (Table 1, Supplementary file 1A, Materials and methods). For each species we included individuals from as wide a geographic distribution as possible to reduce the strength of the signal of divergence from alleles that may be diverged only in certain populations or breeds of a domestic species.

To gain a first impression of levels of divergence between the domesticated and wild animals we plotted the genome-wide distributions of differences in allele frequencies for each domestic and wild species pair (Figure 2, Supplementary file 1B). Although the distributions vary greatly among the species, the cat stands out in having no SNPs that differ by more than ~80% in allele frequency. This may be because only four wildcat genome sequences were available and perhaps also because of hybridization between domestic and wild cats due to the free roaming life-style of many domestic cats (Bradshaw et al., 1999). In fact, cats are sometimes considered to be semi-domesticated (Montague et al., 2014).

Genomic distribution of functional variants

Under neutrality, variants are expected to be equally distributed across the ΔAF spectrum regardless of the functional consequences they may have, *e.g.* if they change amino acids, fall in splice sites or 5'-untranslated regions (UTRs). In contrast, any enrichment of a particular functional category of variants at high ΔAF s is likely due to positive selection (Carneiro et al., 2014). To investigate the extent to which positive selection has driven allelic differentiation between these populations we calculated M-values following Carneiro et al. (2014) (Materials and methods). M-values are an approach to evaluate, when comparing two populations, whether there is an enrichment for any particular category of variant at different allele frequency thresholds of the ΔAF spectrum.

In the majority of comparisons involving historically domesticated species we find at least one category of putatively functional variants to be significantly enriched among bins of highly differentiated SNPs ($\Delta AF \geq 0.7$, χ^2 -analysis), although the strength of this effect varies among species (Figure 3, Supplementary file 5A, Materials and methods). The only exceptions is the cat, possibly due to the fact that few alleles are

segregating at high frequency between wild and domestic cats. Both coding (missense) and non-coding (5'-UTR, 3'-UTR, splice site) variants are significantly enriched across multiple species comparisons, suggesting that both types of variation have contributed towards the domestication process.

We observe a significant enrichment of synonymous variants at high frequency in some comparisons (Figure 3, Supplementary file 5A). This is unexpected under neutral evolution as synonymous variants are not generally considered to have functional effects, although this is not always the case (Lawrie et al., 2013). We suspect that this signal may be explained due to synonymous variants hitchhiking on haplotypes containing selected variants. For example, three of the six genes with synonymous variants at $>0.9 \Delta AF$ in the Asian pig comparison also contain putatively functional variants segregating at high frequency (missense variants in two genes, 3'-UTR variant in one gene).

We do not observe any significant enrichment of functional variants at high frequency in the rats and mink, the two artificially selected lines. This may be due to strong genetic drift that has also occurred in these lines due to small population size, or to that selection has acted on relatively few variants with large effect sizes during the artificial generation of these lines.

We observe only in the rat lines the occurrence of any fixed differences that result in the gain of a premature stop-codon (Supplementary file 5A), suggesting that loss-of-function mutations have not played a major role in domestication, in agreement with the findings of a previous study of rabbit domestication (Carneiro et al., 2014).

Genes shared among domestication events

To identify genes that are highly diverged between historically domesticated animals and their wild relatives we identified highly diverged genes as described above for the rats and mink (Figure 4, Supplementary files 2C-I, Materials and methods). We also ran this analysis comparing two closely related wild species (Asian & European boar) as a control for divergence without domestication (Materials and methods, Supplementary file 2J).

Among the nine domestic-wild species pairs, cats stand out in having only 134 highly diverged genes whereas the others vary between 1,023 and 3,466 (Supplementary file 1C). We observe 3,166 highly diverged between the Asian and European wild boar control comparison. To look for evidence of convergent selection we tested whether there was a significant excess of highly diverged genes shared between comparisons using a resampling approach (Materials and methods). We find a significant excess of gene sharing in 20 out of 36 pairwise comparisons (Supplementary file 5F). Notably, the highly diverged genes in the tame and aggressive rat comparison only show significant overlap with the highly diverged genes in the

tame and aggressive mink lines. However, we also find a significant overlap between highly diverged genes between the comparison between Asian and European boar and four domestic-wild comparisons, suggesting that many of these genes may be prone to rapid divergence and are not unique to the domestication process (Supplementary file 5F). Although some species show a significant excess of sharing in pairwise comparisons, the total amount of sharing across all pairwise comparisons is not significantly greater than expected by chance (Supplementary file 5G, Materials and methods). However, we observe a significant excess of putatively selected genes shared across three species pairs (p -value =0.004), as well as four (p -value =0.001), five ($p < 0.001$) and six (p -value =0.001) domestic-wild species pairs (Supplementary file 5G).

Shared pathways

Whereas individual genes may not have been consistently selected during domestication, certain biological pathways may have been. We therefore tested if biological pathways as defined in KEGG are over represented among the highly differentiated genes found in the nine pairs of domestic and wild species (Supplementary files 3B-J, Materials and methods). Six KEGG pathways are over represented across all nine species comparisons (Supplementary file 5C). This is significantly more sharing than expected by chance ($p < 0.01$, Materials and methods). Three of these pathways are too general to be easily interpretable ('metabolic pathways', 'endocytosis', 'focal adhesion') and one is specific yet hard to interpret ('dilated cardiomyopathy'). These four pathways are also enriched in a comparison between the two non-domesticated yet closely related species, the Asian and European boars (Materials and methods), suggesting that these pathways may be prone to rapid evolution after species divergence (Supplementary file 3K). The remaining two pathways, where divergent genes are enriched in all nine species comparisons but not the two wild boars, are 'axon guidance' and 'ErbB signaling'. Below, we discuss examples of individual genes in these two pathways that are highly diverged in multiple domesticated species and possible phenotypic traits underlying their selection.

Axon guidance

Among five genes that are involved in axonal guidance and found to be highly differentiated in four or more species comparisons are *PLXNA2* and *PLXNC1*, which encode plexin receptors for the semaphorins, proteins that regulate axon guidance and neuronal development (Supplementary file 5D). Because they are located on different chromosomes, linkage cannot explain why both genes are highly differentiated in several domesticated species. *PLXNC1* has been previously reported as a target of selection in pigs (Frantz et al., 2015) and *PLXNA2* has been associated with schizophrenia (Mah et al., 2006) and anxiety in humans (Wray et al., 2007). In cattle, it is located in a QTL for temperament (Gutiérrez-Gil et al., 2008). Four additional plexin genes occur in at least one comparison, as well as 12 semaphorin genes (Supplementary file 5D).

Another group of highly differentiated genes involved in axon guidance is the ephrin-A family of tyrosine kinase receptors. *EPHA5* appears in dog, rat, and sheep comparisons and *EPHA7* in European and Asian pig, rabbit comparisons. *EPHA5* is involved in axon outgrowth during development and synaptic plasticity in the mature brain (Gao et al., 1998). Male mice in which *EphA5* was homozygously inactivated display increased levels of offensive and defensive aggression (Sheleg et al., 2015), a behavior that may be mediated through altered levels of hypothalamic serotonin (Mamiya et al., 2008). Additionally, *EPHA3* is highly differentiated in the dog and rat comparisons, and *EPHA6* and *EPHA4* in the mink and the rat comparisons, respectively (Supplementary file 5E). Thus, domestication has frequently targeted members of the ephrin-A family of neuronal receptors.

ErbB signaling

The ErbB signaling pathway also stands out in all nine species comparisons (Supplementary file 5C). All four genes encoding ErbB receptors are highly differentiated in at least some comparisons: *EGFR* in dogs and sheep, *ERBB2* in Asian and European pigs, *ERBB3* in cats and rabbits, and *ERBB4* in rats. Similarly, all four genes encoding neuregulins, a family of the ErbB receptor ligands, are highly differentiated in several comparisons: *NRG1* in European pigs and goat, *NRG2* in rabbits and rats, *NRG3* in European pigs and rats, and *NRG4* in Asian and European pigs, goats, rats, and sheep. Indeed, all genes that encode ErbB cell-surface receptors and almost all their ligands are highly differentiated in one or more of the domesticated and wild species pairs analyzed (Figure 5). Thus, domestication has recurrently targeted ErbB receptor signaling.

What aspects of the physiology of domesticated animals may have been affected by changes in ErbB signaling? In the hypothalamus, ErbB is involved in the release of gonadotropin-releasing hormone (GnRH) (Clasadonte et al., 2011), a peptide hormone involved in the initiation of the hypothalamic-pituitary-gonadal (HPG) axis (Ojeda and Ma, 1998; Jun Ma et al., 1992; Messina et al., 2016). In rats, selective blockade of ErbB-1 receptors in the hypothalamus delays the onset of female sexual maturity (Ojeda et al., 1990) and the onset of sexual maturity mice is delayed in mice expressing an ErbB-4 receptor lacking the intracellular domain (Prevot et al., 2003). Furthermore, intronic SNPs in the gene encoding the ErbB-2-interacting protein *ERBB2IP* are associated with litter size in pigs (Spötter et al., 2008). One possibility is therefore that changes in ErbB signaling have contributed to the early onset of sexual maturity and reproduction and perhaps in the abolishment of seasonal reproduction seen in many domesticated animals (Sánchez-Villagra et al., 2016; Ojeda et al., 2004; Prevot et al., 2005). In support of this, two other pathways that are also involved in reproduction have similarly highly differentiated genes in seven or eight of the nine species comparisons and are not significantly enriched in the control comparison between European and Asian boar: 'gonadotropin-releasing hormone (GnRH) signaling' and 'progesterone-mediated oocyte maturation' (Supplementary file 5C).

The GnRH signaling pathway is particularly striking, 38 genes in this pathway are highly differentiated in at least two species comparisons. The gene encoding the gonadotropin-releasing hormone receptor, *GNRHR*, is highly diverged in the rats and in the European and Asian pig. Polymorphisms in *GNRHR* are associated the timing of puberty in cattle (Lirón et al., 2011) and the number of eggs laid in chickens (Xu et al., 2007). *GNRHI*, the gene encoding gonadotropin-releasing hormone 1, is highly differentiated between domestic and wild rabbits. Mutations in *GNRHI* cause hypogonadism and delayed or absent puberty in humans (Bouligand et al., 2009) and in domestic goats polymorphisms in *GNRHI* are associated with litter-size (An et al., 2013). Indeed, chemicals that inhibit or stimulate GnRH production are used in domestic animal management to control fertility and regulate behavior (Adams, 2005). It thus seems that variants in genes involved in reproductive phenotypes such as the timing of sexual maturity (Sánchez-Villagra et al., 2016; Boitani and Ciucci, 1995; Schütz et al., 2002), litter size, and the regulation of seasonal reproduction (Trut et al., 2009; Faya et al., 2011; Karlsson et al., 2016; Rosa and Byrant, 2003; Setchell, 1992) have been selected during domestication.

ErbB signaling is also involved in neural crest cell migration, a process that may be delayed in many domesticated species (Wilkins et al., 2014). In particular, ErbB signaling is necessary for neural crest cell pathfinding and pigment pattern formation (Birchmeier, 2009; Budi et al., 2008). Indeed, the gonadotropic neurons in the hypothalamus that respond to ErbB signaling and initiate the HPG axis are themselves derived from neural crest cells and require ErbB signaling for their migration during development (Forni et al., 2011; Whitlock et al., 2003). In mice, mutations in the genes *Nrg1*, *ErbB2*, *ErbB3* and *ErbB4*, which are highly differentiated in several of the domesticated species (Figure 5), result in deficits in neural crest migration (Britsch et al., 1998; Golding et al., 2000).

Convergent selection in the experimental lines

To ask which genes or pathways may have been selected early during domestication rather than during the millennia that followed the initial domestication events, we analyzed the rat and mink lines, which were selected over a short time solely for their behavioral response to humans. We find seven genes that are both highly diverged and carry putatively functional variants that differ dramatically in frequency between the tame and aggressive animals both among the rats and the mink (Supplementary file 4C). This is significantly more than expected by chance (p-value <0.001, Materials and methods). Of these seven genes, two are implicated in defects in the development of tissues derived from the neural crest. One of these genes *CCNG1*, encodes cyclin G, an inhibitor of cell growth. Increased expression of *CCNG1* causes Treacher Collins syndrome, characterized by craniofacial defects (Jones et al., 2008). The other gene, *ECE1*, encodes endothelin-converting enzyme 1, which takes part in the activation of the endothelins, several of which are involved in neurocristopathies, for example piebaldism, a problem with melanocyte development that results in depigmentation of areas of the skin and often a white patch on the forehead. *Ece1*-homozygous knockout

mouse embryos have craniofacial abnormalities and lack epidermal melanocytes (Yanagisawa et al., 1998; Clouthier et al., 1998), features that echo differences observed between the tame and aggressive rat and mink lines (Gulevich et al., 2010). Interestingly, in rabbits, which were domesticated in the Middle Ages and may thus be somewhat similar to the rat and mink in that they are recently domesticated, two variants in the 3'-UTR of *ECE1* differ drastically in frequency between the domesticated and wild animals. Thus, these results lend support to the hypothesis that selection for tameness during early stages of domestication may have driven selection on variants that affect neural crest migration (Wilkins et al., 2014).

DISCUSSION

Several studies of domestication in individual species have suggested it is a highly polygenic and complex trait (Axelsson et al. 2013; Carneiro et al. 2014; Frantz et al. 2015). Our comparative approach, using newly generated whole genome sequences from populations of experimental lines of rats and mink that have been divergently selected for docility towards humans in conjunction with genome sequences from seven historically domesticated cases of animals domestication, reveals genes and pathways that appear to have been recurrently affected by domestication processes.

The functional targets of selection

We find a significant enrichment of both coding and non-coding variation segregating at high allele frequencies between domestic and wild species, suggesting that both types of variation have contributed towards adaptation during domestication. However, we find that non-coding variants, particularly those in the 5'-UTR and splice sites of genes are the most commonly over-represented category of functional variants that differ drastically in allele frequencies between wild and domestic animals, suggesting that these have played a major role during domestication.

In agreement with other studies on domestic animals (Carneiro et al. 2014; Rubin et al. 2010) we find no premature gains of stop-codon variants segregating between historically domesticated animals and their wild sister species, suggesting that loss of function variants have not played a major role in the domestication process.

In contrast to this, we do observe loss of function variants segregating between the tame and aggressive rat lines. We suspect that this difference in the types of variants targeted by selection may be due to the nature of the experimental selection to which the rats were exposed. They have experienced strong founder effects, which must have limited the amount of genetic variation available for selection to act upon. Furthermore, selection has been strong and consistent for one single trait, the behavioral response to an approaching human. Therefore, it is likely that large-effect variants were selected that may have pleiotropic effects that

would be deleterious to fitness in other contexts. Such variants seem to not have been selected during the historical domestication processes. This raises questions about the suitability of these experimental lines as models for historically domesticated animals. However, although the functional categories of variants targeted by selection may differ in the experimental lines, the similarities both in the genes and pathways affected, and in phenotypic changes observed, in these rat and mink lines and in historically domesticated species suggest that the genes affected in the experimentally domesticated lines are involved in pathways affected in most domestication events.

Genes commonly selected during domestication

We find no evidence for a single ‘domestication gene’ which would have been selected across all nine species comparisons. However, we do find a significant excess of convergently selected genes across the species studied (Supplementary file 5G). Several of these genes shared across multiple comparisons have previously been identified as targets of selection in domestic animals, such as the genes *KIT* and *MITF* (Supplementary file 5B). Both these genes are involved in the regulation of the neural crest-derived melanocyte lineage (Hou et al., 2000) and are known to underlie coat-color phenotypes in several domestic animal species (Schmutz et al., 2009; Kijas et al., 2012; Pielberg et al., 2002; Wang et al., 2015; Haase et al., 2013).

KIT is highly differentiated in the dog, European pig, goat and sheep comparisons. It encodes the KIT proto-oncogene receptor tyrosine kinase. Mutations in this gene cause piebaldism and the white spotting pattern in multiple domesticated species (Wright, 2015). *MITF* is highly differentiated in the dog, European pig, goat, mink and sheep comparisons. It encodes a transcription factor involved in the development of neural-crest-derived melanocytes (Nakayama et al., 1998). Mutations in *MITF* cause white-spotting in dogs, (Karlsson et al., 2007), cattle (Hayes et al., 2010) and horses (Hauswirth et al., 2012). In humans, mutations in *MITF* cause Waardenburg syndrome type 2a, characterized by defects in melanocyte and craniofacial development, and a forelock of white hair, similar to the 'white star' pigmentation phenotype seen in many domesticated animals. Notably, mice carrying a mutation that results in aberrant splicing and reduced expression of *Mitf* fail to display anxiety when confined in an enclosed space (Takeda et al., 2014), suggesting that *MITF* may be involved in a behavior of importance during domestication.

Pathways involved in domestication

Six biological pathways appear to be targeted by selection in all nine species comparisons and 20 biological pathways appear affected in eight out of nine species comparisons (Supplementary file 5C). This suggests that although the genes underlying the domestication process vary between species, the striking phenotypic similarities observed across domestic animals (Wilkins et al., 2014; Sanchez-Villagra et al., 2016) may be the

result of recurrent selection on common biological pathways. Notably the ‘Axon guidance’ pathway is shared across all nine species comparisons. This suggests that neurological changes have had an important role in the domestication process, potentially related to the tame behavior of domestic animals. The fact that these pathways are affected also in rats and mink that were experimentally selected for tameness supports the idea that docility towards humans may have been the selected trait.

We also find enrichment in pathways related to reproduction and development (‘GnRH signaling pathway’, ‘Progesterone-mediated oocyte maturation’). For example, selection on genes and pathways involved in the HPG axis seems to have been an essential early step in the domestication process. Genetic variation that influenced the productivity of domestic animals, by reducing the age at which animals reached maturity, increasing their ability to reproduce under stressful conditions, and decreasing inter-birth intervals are likely to have been targeted during both experimental and historical domestication. This may explain why selection on reproduction-related pathways appear to be so widespread in domestic animal species.

Interestingly, the ‘ErbB signaling pathway’ appears to be targeted by selection across all species. This pathway is involved in multiple developmental processes, making it difficult to speculate on the phenotypic traits that may have been targeted during selection. However, it is intriguing that several of the genes which are highly differentiated between domestic animals and their wild sister species (*NRG1*, *ERBB2*, *ERBB3*, *ERBB4*) result in deficits in neural crest development when mutated in mice (Britsch et al., 1998; Golding et al., 2000). Indeed, out of seven genes that are both highly diverged and contain putatively functional variants that differ dramatically in frequency between the tame and aggressive rats and mink, two (*CCNG1*, *ECE1*) are involved in the development of tissues derived from the neural crest (Supplementary file 4C). Thus, selection on genes in the ErbB pathway supports the hypothesis that many aspects of the domestication syndrome can be explained by changes in neural crest migration (Wilkins et al., 2014).

MATERIALS AND METHODS

Sample collection for experimental lines

For the experimentally selected lines rat (*Rattus norvegicus*) and mink (*Neovison vison*) DNA samples were obtained from the Institute of Cytology and Genetics, the Siberian Branch of the Russian Academy of Sciences, Academgorodok, Novosibirsk, in Russia. We obtained DNA from 20 rats that are the product of ~70 generations of selective breeding for tame behavior towards humans and 20 rats from a line selected for aggressive and fearful behaviour towards humans for the same number of generations (hereafter the “tame” and “aggressive” lines, respectively). Both lines were founded at the same time from wild rats from the region surrounding Academgorodok. As a proxy for the wild rats that founded these populations we obtained

DNA samples from ten rats recently captured from the wild in the same region. These animals were bred in captivity for three generations without any deliberate selection (hereafter the “unselected” rat line).

We obtained DNA samples from 20 mink from the tame line and 20 from the aggressive line. Both lines were bred for 15 generations under the same selective breeding scheme as the rat lines. We obtained five DNA samples from a mink line bred for 15 generations under the same captive conditions as the selected lines but without any deliberate selection (hereafter the "unselected" mink line). All three mink lines originate from an unselected founder population. As the lines in both species are maintained at a population size of approximately 100 animals per generation we considered 20 individuals sufficient to sample the diversity present in both lines and identify segregating variants. Further information is in Supplementary file 1A.

Genome sequencing, alignment and SNP calling

Double-indexed paired-end sequencing libraries were generated for all rat and mink samples at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. For each species we pooled all samples to avoid batch effects during sequencing that could confound comparisons between the lines. The libraries were sequenced as 125bp paired-end reads using a combination of the Illumina Genome Analyzer II and High-Seq platforms. The tame and aggressive rats were sequenced to a mean coverage of ~4X per individual and the unselected rats to ~1.3X. The mink were sequenced to a mean coverage of ~6X per individual. Bases were called using the machine-learning algorithm freeIbis (Renaud et al., 2013). Illumina adapters were removed and putative chimeric sequences were flagged as failing quality control using the software leeHom with the default parameters (Renaud et al., 2014). Reads were assigned to their sample of origin using deML with default quality thresholds (Renaud et al., 2015). Reads were aligned to reference genomes using the Burrows-Wheeler Algorithm (BWA) with default parameters (Li and Durbin, 2009). Rat reads were mapped to the reference genome assembly (*Rattus norvegicus*, Rno5). As no mink reference assembly was available mink reads were aligned to the most closely related reference assembly, the ferret (*Mustela putorius fura*, MusPutFur1). Duplicate reads were removed using Picard (<http://broadinstitute.github.io/picard/>). Indel realignment, duplicate removal, SNP and INDEL discovery and genotyping were performed across all samples simultaneously (for each species) using the Genome Analysis Toolkit (GATK) pipeline with standard hard filtering parameters following GATK Best Practices recommendations (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013). SNPs were filtered to remove sites with a Mapping Quality score less than 30 and individual genotype calls supported by less than two reads per sample. Per sample depth of coverage was computed using VCFtools (Danecek et al., 2011). To avoid being misled by sites with high levels of missing data we removed all sites with > 50% missing genotype calls in either the tame or aggressive line.

Genomic data used for domestication meta-analysis

To identify genes and genomic variants involved in the domestication process across species we gathered publicly available whole-genome population data from seven domestic species and extant representatives of the wild species from which they originate (Table 1, Supplementary file 1A). We tried to maximize the number of samples representing each species and to include samples from as wide a geographic range as possible, to increase the likelihood that our allele frequency estimates would be representative of the true species allele frequencies. As all samples come from publicly available data sets the sample sizes can vary substantially between comparisons. For example, for domestic animal genome we have 249 domestic sheep genomes compared to 20 domestic Asian pig genomes and wild animal genomes range from 192 wild rabbits to 4 wild cats (Table 1). Therefore, our power to detect alleles segregating at high frequency between domestic and wild species is expected to vary between comparisons. As with the experimental lines, to reduce inaccurate estimates of allele frequencies for each comparison we removed all sites with > 50% missing genotype calls in either the domestic or wild samples.

Identification of putatively selected genes

For each population comparison for every gene we included all polymorphic sites (SNPs) and substitutions in exons, introns and up to 500bp up/downstream of the coding regions. Incorporating all these sites we calculated the mean ΔAF score for each gene. To identify genes with high levels of divergence we used a bootstrapping approach under the assumption that the majority of genes are not influenced by positive selection. The variance in the mean ΔAF score is expected to be inversely correlated with the total number of sites used to calculate the mean. As a result, the lower the number of informative sites the more likely a gene is to appear highly diverged by chance. Therefore, when identifying highly diverged genes we control for the total number of sites.

For a gene of size x (here size refers to the total number of informative sites not absolute gene length) we consider all genes from the genomic distribution with a similar number ($\pm 20\%$) of informative sites. We combine all observed ΔAF scores of the informative sites from these genes to create a probability distribution of ΔAF scores for genes of size x . Randomly sampling with replacement x times from this probability distribution we calculate a mean ΔAF score for the gene. This sampling process is repeated 1 million times. Using this distribution of mean ΔAF scores we calculate the 99th quantile of mean ΔAF scores for genes with x informative sites. By repeating this process for all observed gene sizes we can create a distribution of expected mean ΔAF for genes that controls for variation in the number of informative sites. We consider all genes with between 1-4000 informative sites. We identify as outliers all genes from the observed data with a mean $\Delta AF >$ the 99th quantile from this bootstrapping approach (Figure 3). The number of genes identified as outliers from this analysis ranges from 178 in cats to 4099 in the rat lines (Supplementary file 1C). To facilitate comparison across species we converted Ensembl gene IDs to their

associated gene symbols (Kinsella et al., 2011; Cunningham et al., 2015). The highly diverged genes for each species comparison are presented in Supplementary files 2A-J.

Highly diverged genes in tameness QTL

Previously, an F2 intercross of the tame and aggressive rat lines was generated and used to identify QTL underlying the phenotypic differences between these lines (Albert et al., 2009). We used LiftOver to convert the QTL coordinates from the rat genome assembly Rno3.4 to Rno5. We identified 4064 genes within these QTL regions. We compared the genes with high mean ΔAF ($>0.99\%$ quantile from bootstrap analysis) in the rat lines (4099) to genes in QTL (4639) identified in an F2 intercross of tame and aggressive rats derived from these lines). We used the software WebGestalt to test for enrichment of biological pathways among these genes (Wang et al., 2013). We used the KEGG ontology and a significance threshold of 0.05 after adjusting for multiple testing with the Benjamini Hochberg procedure. To predict the functional consequences of sequence variants we used Ensembl's Variant Effect Predictor (McLaren et al., 2016).

M-value analysis

To identify functional categories of SNPs that are highly diverged between domestic and wild species pairs we performed an M-value analysis previously used by Carneiro et al., (2014). To calculate M-values we did the following. For each domestic-wild and tame-aggressive population comparison at each polymorphic site we calculated the absolute allele frequency difference (ΔAF) between the two populations. We then divided these into bins based on their ΔAF (0.1-0.19, 0.2-0.29, 0.3-0.39, 0.4-0.49, 0.5-0.59, 0.6-0.69, 0.7-0.79, 0.8-0.89, 0.9-0.99, 1). In all comparisons the majority of SNPs have low ΔAF values. (Supplementary file 1B). To test for enrichment of putatively functional variation among highly divergent sites we further subdivided bins of variants by their functional annotation as defined by Ensembl's Variant Effect Predictor. We considered variants to be functional if they occurred in the 5'-UTR, 3'-UTR, a splice site, result in a missense mutation or the gain of a premature stop codon. We also considered synonymous, intronic and intergenic variants as neutral categories of sites.

For each category of sites we calculated an M-value (\log_2 fold change) of the relative frequency of variants in a given ΔAF bin compared to their frequency across all bins. This was calculated following the approach of Carneiro *et al.*, (2014) (Supplementary file 5A). Briefly, the expected number of variants in each functional category in each bin was calculated as $p(\text{category}) \times n(\text{binX})$, where $p(\text{category})$ is the proportion of a specific variant category across all bins and $n(\text{binX})$ is the total number of variants in a given ΔAF bin. M-values were calculated as the \log_2 fold change of the observed vs the expected variant count. M values >0 indicate that more variants of a given category are present in a ΔAF bin than expected if variants are randomly distributed across bins. As some noise is expected when calculating M-values due to genetic drift

only consider an M-value >0.3 to indicate that a particular category of SNP is enriched. This threshold was determined because it was the maximum M-value observed using intronic and intergenic sites, the majority of which are assumed to be evolving neutrally. We determined statistically significant deviations from the expected values using a χ^2 -analysis (d.f.=1). Plots of the M-value results for each variant category comparison are shown in Figure 3.

Significant sharing of targets of selection across species comparisons

As with other outlier approaches using divergence based statistics (e.g. Fst, iHS), our method to detect putatively selected genes based on high mean ΔAF is unable to discriminate between regions that are highly diverged due to genetic drift or positive selection. Given the strong population bottlenecks associated with the domestication process we expect each individual list of highly diverged genes to contain false positive genes that are highly diverged due to genetic drift. We aim to utilise the strength of the comparative approach to increase our power to identify those highly diverged genes that were targeted by positive selection. While genetic drift acts randomly across the genome, positive selection acts in a locus specific manner. Therefore, if there are genes that are recurrently selected during the domestication process we should observe more sharing of highly diverged genes across species comparisons than expected by chance.

To test this, for each of the nine domestic-wild and tame-aggressive comparisons and a control comparison between European and Asian wild boar (see below), we randomly sampled without replacement the same number of genes that are identified as highly diverged. Genes were sampled from distributions that included all genes from their relevant genomic background with between 1-4000 informative sites, which include polymorphic sites and substitutions (genes that could have been identified as highly diverged). Using these randomly sampled gene lists we calculated the amount of pairwise sharing between each comparison. This process was repeated 1000 times to obtain empirical p-values (Supplementary file 5F).

As well as specific pairwise comparisons we also tested whether the total amount of sharing of highly diverged genes across all domestic-wild and tame-aggressive pairwise comparisons was greater than expected by chance. Using the randomly sampled gene lists we calculated the total amount of sharing across comparisons from pairwise up to nine-way combinations. We repeat this process 1000 times in order to obtain empirical p-values for the probability of the sharing we observe of candidate genes across species (Supplementary file 5G).

Gene pathway enrichment analyses

We performed gene pathway enrichment analyses to identify biological pathways putatively targeted by selection in the artificially selected lines and during the domestication process. To do this we tested each list

of genes identified as highly divergent between the domestic-wild and tame-aggressive pairs for enrichment of KEGG ontology pathways using the software Webgestalt. Over-representation was tested using the hypergeometric test with a significance threshold of 0.05 after adjusting for multiple testing with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). All enriched categories and their constituent genes are presented in Supplementary files 3B-J.

We find multiple KEGG pathways that are enriched across multiple species comparisons. To assess our false discovery rate for each species comparison we randomly sampled the same number of genes that are identified as highly diverged using the same procedure as described above for the gene-sharing resampling analysis. As described, we only resampled genes from the genome-wide background of each species that had the potential to be identified as highly diverged (between 1-4000 informative sites). Using these randomly sampled gene lists we tested for enrichment of KEGG pathways using the same parameters as for the observed data (p-value cut-off = 0.05, Benjamini-Hochberg procedure to adjust for multiple testing). This process was repeated 100 times for each species-comparison. Using the resulting lists of enriched KEGG pathways we tested the number of shared pathways across species comparisons. In 12 out of 100 resamplings we observe one pathway that is significantly enriched across all nine species comparisons. In 11 cases this is the 'Metabolic pathways' and in the remaining case it is the 'focal adhesion' pathway. Therefore, the six pathways that we observe as significantly enriched across all species-comparisons in our data are significantly more than expected by chance ($p < 0.01$).

Identification of highly diverged genes between wild species

By identifying genes that are highly diverged between domestic animals and their wild sister species we aim to find genes that were positively selection during the early domestication process. However, it may be that we also detect genes that tend to evolve rapidly or that are recurrently involved in the speciation process but are not specific to domestication. To investigate this we repeated our analysis to identify highly diverged genes, this time comparing two wild species, the Asian and the European wild boar species. Although the ~500,000 year split time of these species is considerably older than the domestication events we are studying (Giuffra et al., 2000) this comparison provides an opportunity to identify those genes and pathways in our analyses that may frequently diverge during the speciation process but which are not specifically related to domestication. Comparing the Asian and European wild boar species we identify 3,166 genes as highly diverged (>99th quantile threshold)(Supplementary file 2J), which falls within the range of genes identified in the domestic-wild comparisons (Supplementary file 1C). We then tested these genes for enrichment of KEGG pathways. Of the 26 pathways that are enriched across at least eight of the nine domestic-wild comparisons, 14 are enriched in the wild boar comparison (Supplementary file 3K). They include pathways related to immune function ('Toll-like receptor signaling' pathway, 'Leukocyte transendothelial migration') and metabolism ('Lysine degradation', 'Metabolic pathways'). Notably, neither the 'Axon guidance' nor the

'ErbB signaling pathways', which are enriched across all nine domestic-wild species comparisons, nor any hormonal pathways ('GnRH signaling pathway', 'Progesterone mediated oocyte maturation') are enriched in the wild boar comparison. These results suggest that we are able to detect pathways that appear to be specifically and recurrently involved in the domestication process.

Putatively selected variants in the experimental lines

Genomic regions with signals of selection often contain multiple genes due to the process of hitchhiking, which raises the frequency of variants linked to the selected allele. This is of particular interest for the experimentally selected lines of rat and mink, where selection for behavior has been extremely strong and consistent (~10% of each generation selected to breed based on their behavioral response to a human, resulting in a selection coefficient of ~0.1) and as a result where hitchhiking effects may be strong. We tried to mitigate this problem by taking advantage of the whole-genome sequence data to look directly for putatively functional variants that are fixed for alternative alleles or segregating at high frequency between the lines.

To identify the putatively causal variants targeted by selection in the experimental lines we first selected a comparable number of sites segregating at high frequency in the rat and mink lines. The tame and aggressive rats have 83,459 sites fixed for alternative alleles (83,459). To have a comparable number of sites segregating at high frequency for the mink we set a cut-off in the ΔAF spectrum of all sites above 0.45 ΔA , which resulted in 69,510 sites (Supplementary file 4B). We considered as putatively functional any variants that reside in 3'-UTRs, 5'-UTRs, predicted splice sites, cause frameshift or missense mutations or the gain of a premature stop codon. These lists of putatively selected variants was further refined to only include variants in genes that were identified as highly diverged based on their mean ΔAF . This resulted in 207 and 221 genes that were both highly diverged and contain putatively functional variants segregating at high frequency in the mink and rat lines respectively. Comparing these lists we find seven genes that are candidate targets of selection in both the rat and mink lines (Supplementary file 4C). To test if this overlap is significantly more than expected by chance we performed permutation testing. We randomly sampled an equivalent number of genes (207 and 221) from each line from the genomic background, excluding genes with >4000 SNPs in accordance with our method to identify highly diverged genes, and counted the observed overlap. This was repeated 1000 times. We find that the observed overlap of seven genes is significantly greater than expected by chance (p-value <0.001). The mean overlap from permutation testing is 1.1 and the maximum is five genes.

Ethics statement

Care of the experimental rat and mink lines used in this study was in accordance with institutional guidelines.

Acknowledgments

We are grateful to the NextGen project for allowing early access to their data. This work was supported by the State budget project (no. 53.2.3) in Russia and by the Max Planck Society in Germany.

Figure Legends

Figure 1. PCA plots of rat and mink samples. PCA calculated from whole genome data for tame, aggressive and unselected rat (A) and mink (B) lines. Samples are color-coded according to line.

Figure 2. Δ allele frequency spectra for domestic-wild comparisons. Δ allele frequency spectra (Δ AF) divided into bins for domestic-wild comparisons. The title of each plot refers to the domesticated species in the comparison. X-axis: Non-cumulative bins of the Δ AF. Y-axis: Number of SNPs.

Figure 3. Proportion of functionally annotated SNPs in bins of the delta allele frequency spectrum. Colored lines denote M-values (log₂-fold changes) of the relative frequencies of SNPs at different categories of functional sites according to Δ AF bins (x axis). M-values were calculated by comparing the frequency of SNPs in a given annotation category in a specific bin to the corresponding frequency across all bins. Total functional sites represents the combined total of 3'-UTR, 5'-UTR, missense, splice sites and gain of stop codon sites. M-values > 0 indicate an increase in the proportion of that category of variant in the Δ AF bin relative to neutral expectations. Bins with significant (χ^2 -analysis, $p \leq 0.05$) M-values > 0.3 are labelled with an asterisk in the color of the relevant species comparison. X-axis: Non-cumulative bins of the Δ AF. Y-axis: M-value.

Figure 4. Identification of highly diverged genes. For each species comparison we calculated the mean Δ AF of all SNPs (including substitutions) in and 500bp up/downstream of each gene. Each point is the mean Δ AF for a single gene. The 99th quantile of the expected mean Δ AF obtained from a resampling approach is plotted as a red line. Genes above the 99th quantile threshold were considered highly diverged. X-axis: Number of SNPs used to calculate the mean Δ AF for each gene. Y-Axis: Mean Δ AF score.

Figure 5. Convergent selection on receptors in the ErbB signaling pathway. Diagram representing proteins involved in ErbB receptor signaling. The ErbB receptors and proteins directly interacting with them in the KEGG ErbB signalling pathway are shown. Proteins encoded by genes with evidence of selection in at least two species comparisons are highlighted in blue.

Figure 1. PCA plots of rat and mink samples.

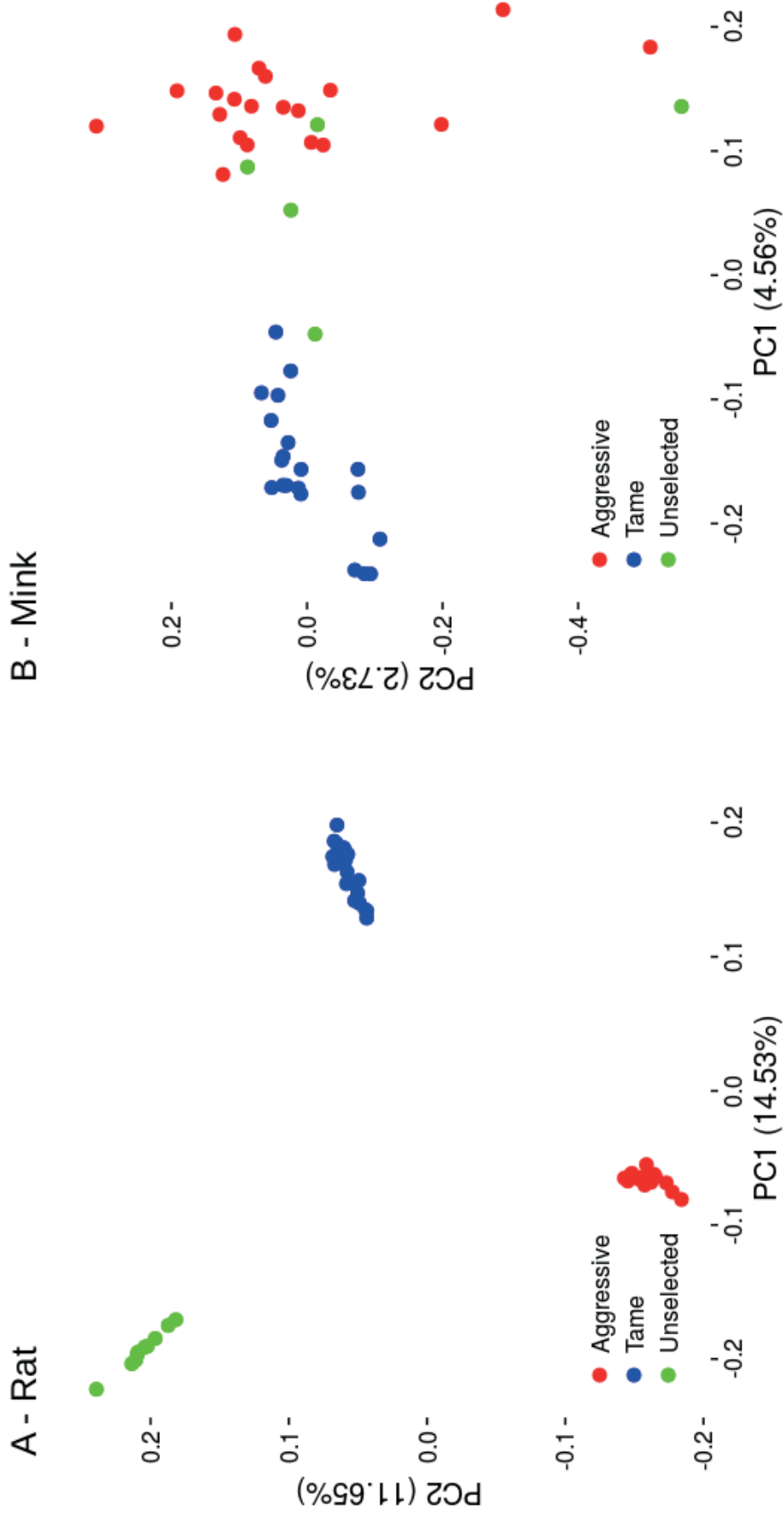


Figure 2. Δ allele frequency spectra for domestic-wild comparisons.

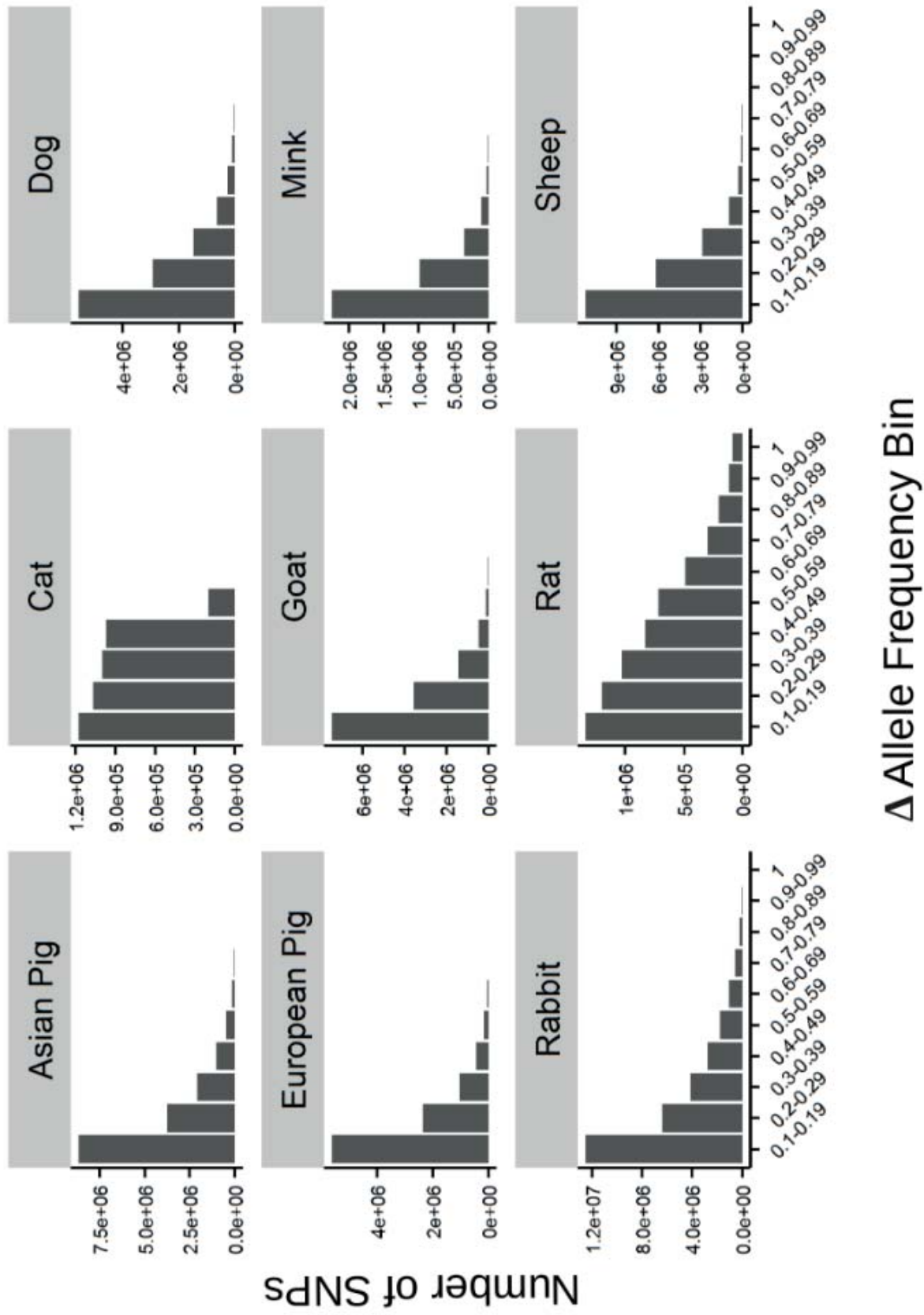


Figure 3. Proportion of functionally annotated SNPs in bins of the Δ allele frequency spectrum.

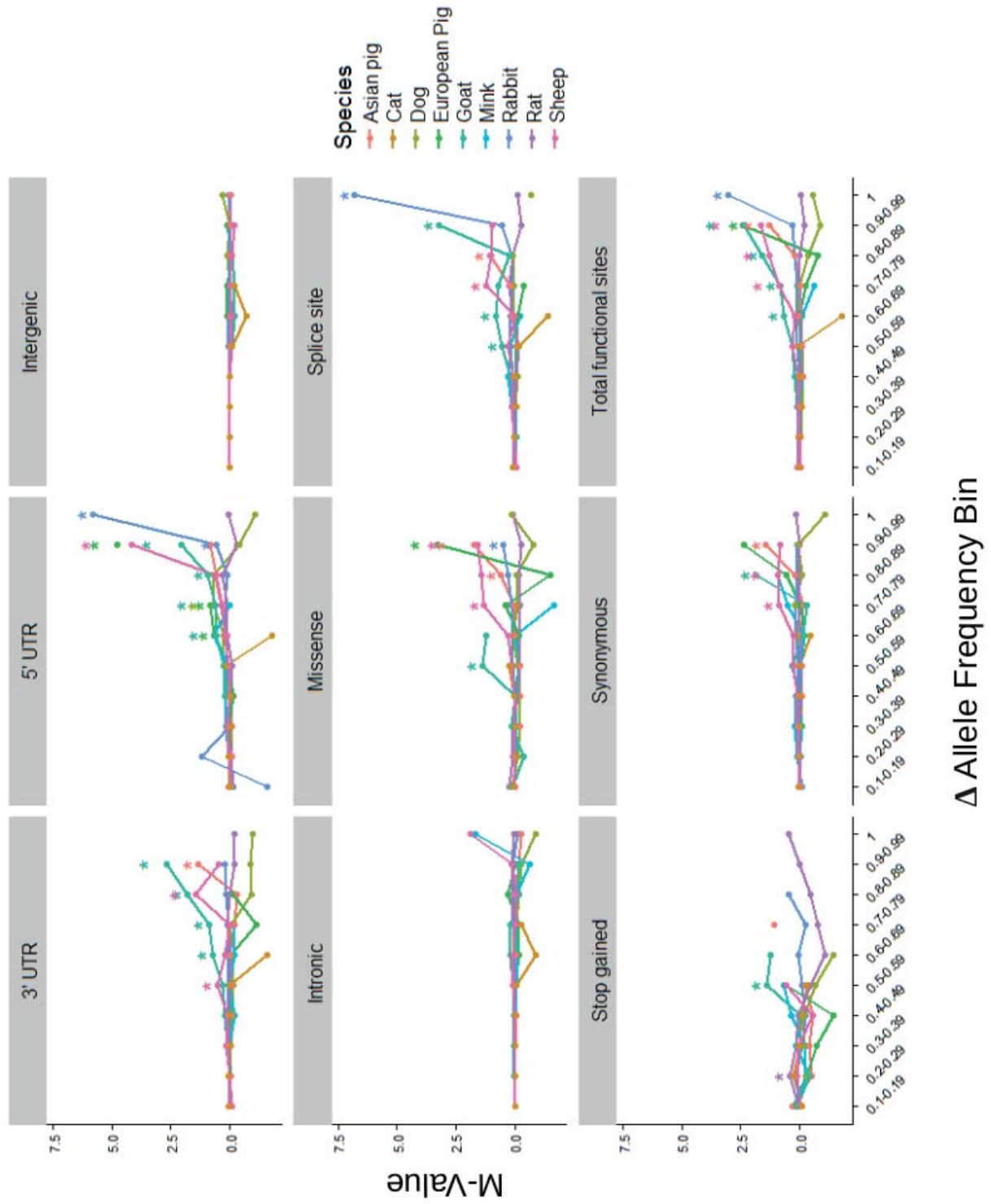


Figure 4. Identification of highly diverged genes.

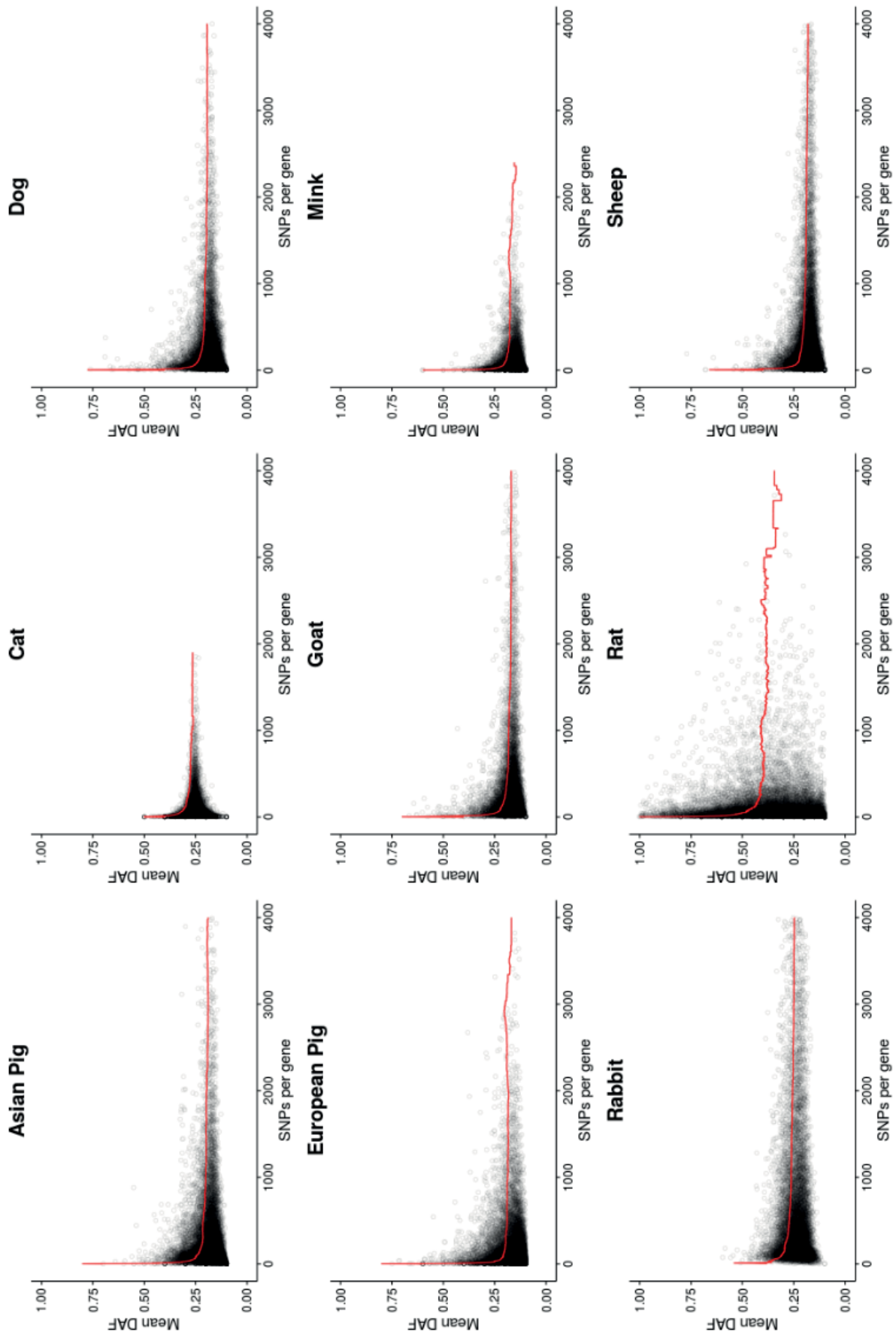


Figure 5. Convergent selection on receptors in the ErbB signaling pathway.

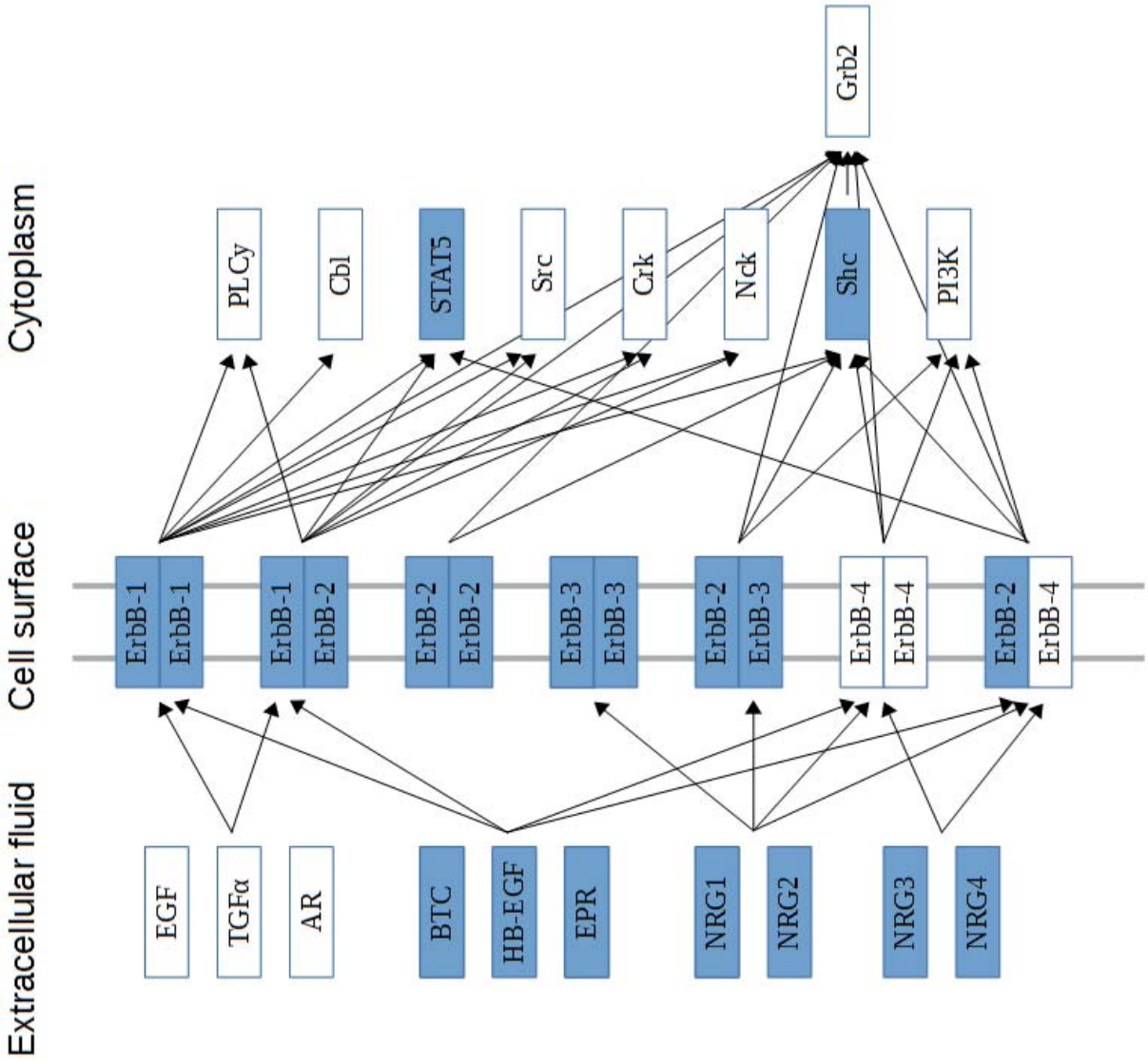


Table 1. Sample Information. Number of samples per species, origin of samples and the reference genome assembly used for each species.

Species pairs	Number of domesticated/tame samples	Number of wild/aggressive samples	Estimated divergence time	Sample origins (references)	Reference genome
Domestic Asian pig – Wild Asian boar	20	11	~10,000 years ^a	Bosse et al., 2014, Frantz et al., 2015	susScr3
Domestic European pig – Wild European boar	45	27	~10,000 years ^a	Bosse et al., 2014, Frantz et al., 2015	susScr3
Domestic cat – Wild cat	27	4	~9,500 years ^b	Montague et al., 2014	FelCat5
Dog - Wolf	67	7	~9,000-33,000 years ^c	Bai et al., 2014	CanFam3
Domestic rabbit - Wild rabbit	100	192	~1,400 years ^d	Carneiro et al., 2014	OryCun2.0
Domestic goat - Wild goat	190	21	~11,000 years ^e	Nextgen Project	Caeg1
Domestic sheep - Wild sheep	249	19	~11,000 years ^f	Nextgen Project	Oori1
Tame mink - Aggressive mink	20	20	15 generations	Generated for this study	MusPutFur1
Tame rat - Aggressive rat	20	20	~70 generations	Generated for this study	Rno5
Wild Asian boar- European boar	11	27	~500,000 years ^g	Bosse et al., 2014, Frantz et al., 2015	SusScr3

^a Bosse et al., 2014^b Vigne et al., 2004^c Skoglund et al., 2011^d Clutton-Brock, 1999^e Zeder and Hesse, 2000^f Zeder, 2008^g Giuffra et al., 2000

Additional Supplementary Files 1-5 are found in USB stick accompanying this thesis.

References

- Adams T. E., 2005 Using gonadotropin-releasing hormone (GnRH) and GnRH analogs to modulate testis function and enhance the productivity of domestic animals. *Anim. Reprod. Sci.* **88**: 127–139.
- Albert F. W., Carlborg O., Plyusnina I., Besnier F., Hedwig D., Lautenschläger S., Lorenz D., McIntosh J., Neumann C., Richter H., Zeising C., Kozhemyakina R., Shchepina O., Kratzsch J., Trut L., Teupser D., Thierry J., Schöneberg T., Andersson L., Pääbo S., 2009 Genetic architecture of tameness in a rat model of animal domestication. *Genetics* **182**: 541–54.
- Albert F. W., Somel M., Carneiro M., Aximu-Petri A., Halbwax M., Thalmann O., Blanco-Aguilar J. A., Plyusnina I. Z., Trut L., Villafuerte R., Ferrand N., Kaiser S., Jensen P., Pääbo S., 2012 A comparison of brain gene expression levels in domesticated and wild animals. *PLoS Genet.* **8**: e1002962.
- An X., Hou J., Zhao H., Li G., Bai L., Peng J., 2013 Polymorphism identification in goat GNRH1 and GDF9 genes and their association analysis with litter size. *Animal genetics.* **44**: 234-8
- Auwera G.A., Carneiro M.O., Hartl C., Poplin R., del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J. and Banks E., 2013 From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics.* 11-10.
- Axelsson E., Ratnakumar A., Arendt M.-L., Maqbool K., Webster M. T., Perloski M., Liberg O., Arnemo J. M., Hedhammar A., Lindblad-Toh K., 2013 The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**: 360–4.
- Bai B., Zhao W.-M., Tang B.-X., Wang Y.-Q., Wang L., Zhang Z., Yang H.-C., Liu Y.-H., Zhu J.-W., Irwin D. M., Wang G.-D., Zhang Y.-P., 2015 DoGSD: the dog and wolf genome SNP database. *Nucleic Acids Res.* **43**: D777-83.
- Belyaev D., 1979 Destabilizing selection as a factor in domestication. *J. Hered.* **70**:301-8
- Benjamini Y., Hochberg Y., 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological).* **1**:289-300
- Birchmeier C., 2009 ErbB receptors and the development of the nervous system. *Exp. Cell Res.* **315**: 611-8.
- Boitani L., Ciucci P., 1995 Comparative social ecology of feral dogs and wolves. *Ethol. Ecol. Evol.* **7**: 49-72.
- Bosse M., Megens H.-J., Madsen O., Frantz L. A. F., Paudel Y., Crooijmans R. P. M. A., Groenen M. A. M., 2014 Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent *Sus scrofa* populations. *Mol. Ecol.* **23**: 4089–102.
- Bouligand J., Ghervan C., Tello J. A., Brailly-Tabard S., Salenave S., Chanson P., Lombès M., Millar R. P., Guiochon-Mantel A., Young J., 2009 Isolated Familial Hypogonadotropic Hypogonadism and a *GNRH1* Mutation. *N. Engl. J. Med.* **360**: 2742–2748.

- Bradshaw J. W. ., Horsfield G. ., Allen J. ., Robinson I. ., 1999 Feral cats: their role in the population dynamics of *Felis catus*. *Appl. Anim. Behav. Sci.* **65**: 273–283.
- Britsch S., Li L., Kirchhoff S., Theuring F., Brinkmann V., Birchmeier C., Riethmacher D., 1998 The ErbB2 and ErbB3 receptors and their ligand, neuregulin-1, are essential for development of the sympathetic nervous system. *Genes Dev.* **12**: 1825–1836.
- Budi E. H., Patterson L. B., Parichy D. M., 2008 Embryonic requirements for ErbB signaling in neural crest development and adult pigment pattern formation. *Development* **135**: 2603–2614.
- Carneiro M., Rubin C.-J., Palma F. Di, Albert F. W., Alföldi J., Barrio A. M., Pielberg G., Rafati N., Sayyab S., Turner-Maier J., Younis S., Afonso S., Aken B., Alves J. M., Barrell D., Bolet G., Boucher S., Burbano H. A., Campos R., Chang J. L., Duranthon V., Fontanesi L., Garreau H., Heiman D., Johnson J., Mage R. G., Peng Z., Queney G., Rogel-Gaillard C., Ruffier M., Searle S., Villafuerte R., Xiong A., Young S., Forsberg-Nilsson K., Good J. M., Lander E. S., Ferrand N., Lindblad-Toh K., Andersson L., 2014 Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* **345**: 1074–9.
- Clasadonte J., Poulain P., Hanchate N. K., Corfas G., Ojeda S. R., Prevot V., 2011 Prostaglandin E2 release from astrocytes triggers gonadotropin-releasing hormone (GnRH) neuron firing via EP2 receptor activation. *Proc. Natl. Acad. Sci.* **108**: 16104–16109.
- Clouthier, D.E., Hosoda, K., Richardson, J.A., Williams, S.C., Yanagisawa, H., Kuwaki, T., Kumada, M., Hammer, R.E. and Yanagisawa, M., 1998 Cranial and cardiac neural crest defects in endothelin-A receptor-deficient mice. *Development* **125**: 813–824.
- Clutton-Brock J., 1999 *A natural history of domesticated mammals*. Cambridge University Press.
- Cunningham F., Amode M. R., Barrell D., Beal K., Billis K., Brent S., Carvalho-Silva D., Clapham P., Coates G., Fitzgerald S., Gil L., Girón C. G., Gordon L., Hourlier T., Hunt S. E., Janacek S. H., Johnson N., Juettemann T., Kähäri A. K., Keenan S., Martin F. J., Maurel T., McLaren W., Murphy D. N., Nag R., Overduin B., Parker A., Patricio M., Perry E., Pignatelli M., Riat H. S., Sheppard D., Taylor K., Thormann A., Vullo A., Wilder S. P., Zadissa A., Aken B. L., Birney E., Harrow J., Kinsella R., Muffato M., Ruffier M., Searle S. M. J., Spudich G., Trevanion S. J., Yates A., Zerbino D. R., Flicek P., 2015 Ensembl 2015. *Nucleic Acids Res.* **43**: D662-9.
- Danecek P., Auton A., Abecasis G., Albers C. A., Banks E., DePristo M. A., Handsaker R. E., Lunter G., Marth G. T., Sherry S. T., McVean G., Durbin R., 2011 The variant call format and VCFtools. *Bioinformatics* **27**: 2156–8.
- DePristo M. A., Banks E., Poplin R., Garimella K. V, Maguire J. R., Hartl C., Philippakis A. A., Angel G. del, Rivas M. A., Hanna M., McKenna A., Fennell T. J., Kernytsky A. M., Sivachenko A. Y., Cibulskis

- K., Gabriel S. B., Altshuler D., Daly M. J., 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**: 491–498.
- Faya M., Carranza A., Priotto M., Abeya M., Diaz J. D., Gobello C., 2011 Domestic queens under natural temperate photoperiod do not manifest seasonal anestrus. *Anim. Reprod. Sci.* **129**: 78–81.
- Forni P., Taylor-Burds C., Melvin V., 2011 Neural Crest and Ectodermal Cells Intermix in the Nasal Placode to Give Rise to GnRH-1 Neurons, Sensory Neurons, and Olfactory Ensheathing Cells. *Journal of Neuroscience.* **31**: 6915-27.
- Frantz L. A. F., Schraiber J. G., Madsen O., Megens H.-J., Cagan A., Bosse M., Paudel Y., Crooijmans R. P. M. A., Larson G., Groenen M. A. M., 2015 Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nat. Genet.* **47**: 1141–1148.
- Gao W., Shinsky N., Armanini M., Moran P., 1998 Regulation of hippocampal synaptic plasticity by the tyrosine kinase receptor, REK7/EphA5, and its ligand, AL-1/Ephrin-A5. *Molecular and Cellular Neuroscience.* **11**:247-59.
- Golding J., Trainor P., Krumlauf R., Gassmann M., 2000 Defects in pathfinding by cranial neural crest cells in mice lacking the neuregulin receptor ErbB4. *Nat. Cell Biol.* **2**:103-9.
- Gulevich R. G., Plyusnina I. Z., Prasolova L. A., Oskina I. N., Trut L. N., 2010 White spotting in Norway rats selected for tame behavior. *J. Zool.* **280**: 264–270.
- Gutiérrez-Gil B., Ball N., Burton D., Haskell M., 2008 Identification of quantitative trait loci affecting cattle temperament. *Journal of Heredity.* **99**:629-38.
- Haase B., Signer-Hasler H., Binns M. M., Obexer-Ruff G., Hauswirth R., Bellone R. R., Burger D., Rieder S., Wade C. M. and Leeb T., 2013 Accumulating Mutations in Series of Haplotypes at the KIT and MITF Loci Are Major Determinants of White Markings in Franches-Montagnes Horses (YG Shellman, Ed.). *PLoS One* **8**: e75071.
- Hammer K., 1984 Das domestikationssyndrom. *The Cultivated.* **32**:11-34.
- Hauswirth R., Haase B., Blatter M., Brooks S., 2012 Mutations in MITF and PAX3 cause “splashed white” and other white spotting phenotypes in horses. *PLoS Genet.* **8**:e1002653
- Hayes B., Pryce J., Chamberlain A., Bowman P., 2010 Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model. *PLoS Genet.* **6**: e1001139.
- Hou L., Panthier J. J., Arnheiter H., 2000 Signaling and transcriptional regulation in the neural crest-derived melanocyte lineage: interactions between KIT and MITF. *Development* **127**: 5379-89.

- Jones N. C., Lynn M. L., Gaudenz K., Sakai D., Aoto K., Rey J.-P., Glynn E. F., Ellington L., Du C., Dixon J., Dixon M. J., Trainor P. A., 2008 Prevention of the neurocristopathy Treacher Collins syndrome through inhibition of p53 function. *Nat. Med.* **14**: 125–33.
- Jun Ma Y., Junier M.-P., Costa M. E., Ojeda S. R., 1992 Transforming growth factor- α gene expression in the hypothalamus is developmentally regulated and linked to sexual maturation. *Neuron* **9**: 657–670.
- Kanehisa M., Goto S., Kawashima S., Okuno Y., Hattori M., 2004 The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**: D277–D280.
- Kanehisa M., Sato Y., Kawashima M., Furumichi M., Tanabe M., 2016 KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**: D457–62.
- Karlsson E., Baranowska I., Wade C., 2007 Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature genetics.* **39**:1321–8.
- Karlsson A.-C., Fallahshahroudi A., Johnsen H., Hagenblad J., Wright D., Andersson L., Jensen P., 2016 A domestication related mutation in the thyroid stimulating hormone receptor gene (TSHR) modulates photoperiodic response and reproduction in chickens. *Gen. Comp. Endocrinol.* **228**: 69–78.
- Kijas J. W., Lenstra J. A., Hayes B., Boitard S., Neto L. R. P., San Cristobal M., Servin B., McCulloch R., Whan V., Gietzen K. and Paiva S., 2012 Genome-Wide Analysis of the World’s Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PLoS Biol.* **10**: e1001258.
- Kinsella R. J., Kähäri A., Haider S., Zamora J., Proctor G., Spudich G., Almeida-King J., Staines D., Derwent P., Kerhornou A., Kersey P., Flicek P., 2011 Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database (Oxford).* **2011**: bar030.
- Kishimoto T., Radulovic J., Radulovic M., Lin C., 2000 Deletion of *chr2* reveals an anxiolytic role for corticotropin-releasing hormone receptor-2. *Nature genetics.* **24**: 415–19
- Larson G., Piperno D. R., Allaby R. G., Purugganan M. D., Andersson L., Arroyo-Kalin M., Barton L., Climer Vigueira C., Denham T., Dobney K., Doust A. N., Gepts P., Gilbert M. T. P., Gremillion K. J., Lucas L., Lukens L., Marshall F. B., Olsen K. M., Pires J. C., Richerson P. J., Rubio de Casas R., Sanjur O. I., Thomas M. G., Fuller D. Q., 2014 Current perspectives and the future of domestication studies. *Proc. Natl. Acad. Sci.* **111**: 6139–6146.
- Lawrie D. S., Messer P. W., Hershberg R., Petrov D. A., 2013 Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet.* **9**: e1003527.
- Li H., Durbin R., 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–60.

- Lind P. a, Luciano M., Horan M. a, Marioni R. E., Wright M. J., Bates T. C., Rabbitt P., Harris S. E., Davidson Y., Deary I. J., Gibbons L., Pickles A., Ollier W., Pendleton N., Price J. F., Payton A., Martin N. G., 2009 No association between Cholinergic Muscarinic Receptor 2 (CHRM2) genetic variation and cognitive abilities in three independent samples. *Behav. Genet.* **39**: 513–23.
- Lirón J., Prando A., Ripoli M., 2011 Characterization and validation of bovine Gonadotropin releasing hormone receptor (GNRHR) polymorphisms. *Research in veterinary science.* **91**: 391-396.
- Mah S., Nelson M., Delisi L., Reneland R., 2006 Identification of the semaphorin receptor PLXNA2 as a candidate for susceptibility to schizophrenia. *Molecular psychiatry.* **11**: 471-8.
- Mamiya P., Hennesy Z., Zhou R., Wagner G., 2008 Changes in attack behavior and activity in EphA5 knockout mice. *Brain Research.* **1205**: 91-9.
- McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M. A., 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**: 1297–303.
- McLaren W., Gil L., Hunt S. E., Riat H. S., Ritchie G. R., Thormann A., Flicek P. and Cunningham, F., 2016 The Ensembl Variant Effect Predictor. *Genome Biol.* **17**: 122.
- Messina A., Langlet F., Chachlaki K., Roa J., Rasika S., Jouy N., Gallet S., Gaytan F., Parkash J., Tena-Sempere M., Giacobini P., Prevot V., 2016 A microRNA switch regulates the rise in hypothalamic GnRH production before puberty. *Nat. Neurosci.* **19**: 835–844.
- Montague M. J., Li G., Gandolfi B., Khan R., Aken B. L., Searle S. M. J., Minx P., Hillier L. W., Koboldt D. C., Davis B. W., Driscoll C. A., Barr C. S., Blackstone K., Quilez J., Lorente-Galdos B., Marques-Bonet T., Alkan C., Thomas G. W. C., Hahn M. W., Menotti-Raymond M., O'Brien S. J., Wilson R. K., Lyons L. A., Murphy W. J., Warren W. C., 2014 Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc. Natl. Acad. Sci.* **111**: 17230–17235.
- Nakayama A., Nguyen M.-T. T., Chen C. C., Opdecamp K., Hodgkinson C. A., Arnheiter H., 1998 Mutations in microphthalmia, the mouse homolog of the human deafness gene MITF, affect neuroepithelial and neural crest-derived melanocytes differently. *Mech. Dev.* **70**: 155–166.
- Naumenko E. V., Popova N. K., Nikulina E. M., Dygalo N. N., Shishkina G. T., Borodin P. M., Markel A. L., 1989 Behavior, adrenocortical activity, and brain monoamines in Norway rats selected for reduced aggressiveness towards man. *Pharmacol. Biochem. Behav.* **33**: 85–91.
- Ojeda S. R., Urbanski H. F., Costa M. E., Hill D. F., Moholt-Siebert M., 1990 Involvement of transforming growth factor alpha in the release of luteinizing hormone-releasing hormone from the developing female hypothalamus. *Proc. Natl. Acad. Sci. U. S. A.* **87**: 9698–702.

- Ojeda S., Ma Y., 1998 Epidermal growth factor tyrosine kinase receptors and the neuroendocrine control of mammalian puberty. *Mol. Cell. Endocrinol.* **140**: 101-6.
- Ojeda S. R., Prevot V., Heger S., Lomniczi A., Dziedzic B., Mungenast A., 2004 The Neurobiology of Female Puberty. *Horm. Res. Paediatr.* **60**: 15–20.
- Patterson N., Price A. L. and Reich D., 2006. Population structure and eigenanalysis. *PLoS genet.* **2**: e190.
- Pielberg G., Olsson C., Syvänen A.-C., Andersson L., 2002 Unexpectedly High Allelic Diversity at the KIT Locus Causing Dominant White Color in the Domestic Pig. *Genetics* **160**: 305-311.
- Prevot V., Rio C., Cho G. J., Lomniczi A., Heger S., Neville C. M., Rosenthal N. A., Ojeda S., Corfas G., 2003 Normal female sexual development requires neuregulin-erbB receptor signaling in hypothalamic astrocytes. *J. Neurosci.* **23**: 230–239.
- Prevot V., Lomniczi A., Corfas G., Ojeda S. R., 2005 erbB-1 and erbB-4 Receptors Act in Concert to Facilitate Female Sexual Development and Mature Reproductive Function. *Endocrinology* **146**: 1465–1472.
- Price A. L., Price A. L., Patterson N. J., Patterson N. J., Plenge R. M., Plenge R. M., Weinblatt M. E., Weinblatt M. E., Shadick N. a, Shadick N. a, Reich D., Reich D., 2006 Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**: 904–9.
- Renaud G., Kircher M., Stenzel U., Kelso J., 2013 freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. *Bioinformatics* **29**: 1208–9.
- Renaud G., Stenzel U., Kelso J., 2014 leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Res.* **42**: e141.
- Renaud G., Stenzel U., Maricic T., Wiebe V., Kelso J., 2015 deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics* **31**: 770–2.
- Rosa H., Bryant M., 2003 Seasonality of reproduction in sheep. *Small Rumin. Res.* **48**: 155-71
- Rubin C.-J., Zody M. C., Eriksson J., Meadows J. R. S., Sherwood E., Webster M. T., Jiang L., Ingman M., Sharpe T., Ka S., Hallböök F., Besnier F., Carlborg Ö., Bed'hom B., Tixier-Boichard M., Jensen P., Siegel P., Lindblad-Toh K., Andersson L., 2010 Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**: 587–591.
- Sánchez-Villagra M. R., Geiger M., Schneider R. A., 2016 The taming of the neural crest: a developmental perspective on the origins of morphological covariation in domesticated mammals. *Royal Society Open Science.* **3**: 160107
- Schmutz S. M., Berryere T. G., Dreger D. L., 2009 MITF and White Spotting in Dogs: A Population Study. *J. Hered.* **100**: S66–S74.

- Schütz K., Kerje S., Carlborg Ö., Jacobsson L., Andersson L., Jensen P., 2002 QTL Analysis of a Red Junglefowl \times White Leghorn Intercross Reveals Trade-Off in Resource Allocation between Behavior and Production Traits. *Behav. Genet.* **32**: 423-33
- Setchell B., 1992 Domestication and reproduction. *Anim. Reprod. Sci.* **28**: 195-202
- Sheleg M., Yochum C., Richardson J., 2015 Ephrin-A5 regulates inter-male aggression in mice. *Behavioural brain research.* **286**: 300-7
- Shen Q., Lal R., Luellen B. A., Earnheart J. C., Andrews A. M., Luscher B., 2010 γ -Aminobutyric Acid-Type A Receptor Deficits Cause Hypothalamic-Pituitary-Adrenal Axis Hyperactivity and Antidepressant Drug Sensitivity Reminiscent of Melancholic Forms of Depression. *Biol. Psychiatry* **68**: 512–520.
- Shikhevich S. G., Os'kina I. N., Plyusnina I. Z., 2003 Responses of the Hypophyseal-Adrenal System to Stress and Immune Stimuli in Gray Rats Selected for Behavior. *Neurosci. Behav. Physiol.* **33**: 861–866.
- Skoglund P., Götherström A., Jakobsson M., 2011 Estimation of population divergence times from non-overlapping genomic sequences: examples from dogs and wolves. *Mol. Biol. Evol.* **28**: 1505–17.
- Spötter A., Hamann H., Müller S., Distl O., 2008 Effect of Polymorphisms in Four Candidate Genes for Fertility on Litter Size in a German Pig Line. *Reprod. Domest. Anim.* **45**: 579–584.
- Takeda K., Hozumi H., Nakai K., Yoshizawa M., Satoh H., Yamamoto H., Shibahara S., 2014 Insertion of long interspersed element-1 in the *Mitf* gene is associated with altered neurobehavior of the black-eyed white *Mitf*(mi-bw) mouse. *Genes Cells* **19**: 126–40.
- Trut L. N., Plyusnina I. Z., Oskina I. N., 2004 An Experiment on Fox Domestication and Debatable Issues of Evolution of the Dog. *Russ. J. Genet.* **40**: 644–655.
- Trut L., Oskina I., Kharlamova A., 2009 Animal evolution during domestication: the domesticated fox as a model. *BioEssays* **31**: 349–360.
- Vigne J.-D., Guilaine J., Debue K., Haye L., Gérard P., 2004 Early Taming of the Cat in Cyprus. *Science* (80-). **304**.
- Wang J., Duncan D., Shi Z., Zhang B., 2013 WEB-based GENE SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41**: W77-83.
- Wang C., Wang H., Zhang Y., Tang Z., Li K., Liu B., 2015 Genome-wide analysis reveals artificial selection on coat colour and reproductive traits in Chinese domestic pigs. *Mol. Ecol. Resour.* **15**: 414–424.
- Whitlock K., Wolf C., Boyce M., 2003 Gonadotropin-releasing hormone (GnRH) cells arise from cranial neural crest and adenohypophyseal regions of the neural plate in the zebrafish, *Danio rerio*. *Dev. Biol.* **257**: 140-52

- Wilkins A. S., Wrangham R. W., Fitch W. T., 2014 The “Domestication Syndrome” in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics. *Genetics* **197**: 795–808.
- Wray N., James M., Mah S., 2007 Anxiety and comorbid measures associated with PLXNA2. *Arch.* **64**: 318-26
- Wright D., 2015. The Genetic Architecture of Domestication in Animals. *Bioinformatics and biology insights.* **9**: 11.
- Xu W., Li H., Yan M., Tang Q., Chen K., 2007 Associations of gonadotropin-releasing hormone receptor (GnRHR) and neuropeptide Y (NPY) genes’ polymorphisms with egg-laying traits in Wenchang chicken. *Agric. Sci. in China.* **6**: 499-504
- Yanagisawa H., Yanagisawa M., Kapur R., 1998 Dual genetic pathways of endothelin-mediated intercellular signaling revealed by targeted disruption of endothelin converting enzyme-1 gene. *Development Cambridge.* **125**: 825-836
- Ye G.-L., Baker K. B., Mason S. M., Zhang W., Kirkpatrick L., Lanthorn T. H., Savelieva K. V., 2010 GABAA Receptor $\alpha 1$ Subunit (Gabra1) Knockout Mice: Review and New Results. *Transgenic and mutant tools to model brain disorders.* 65-90.
- Zeder M. A., Hesse B., 2000 The Initial Domestication of Goats (*Capra hircus*) in the Zagros Mountains 10,000 Years Ago. *Science.* **287**: 2254-2257.
- Zeder M. A., 2008 Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc. Natl. Acad. Sci. U. S. A.* **105**: 11597–604.

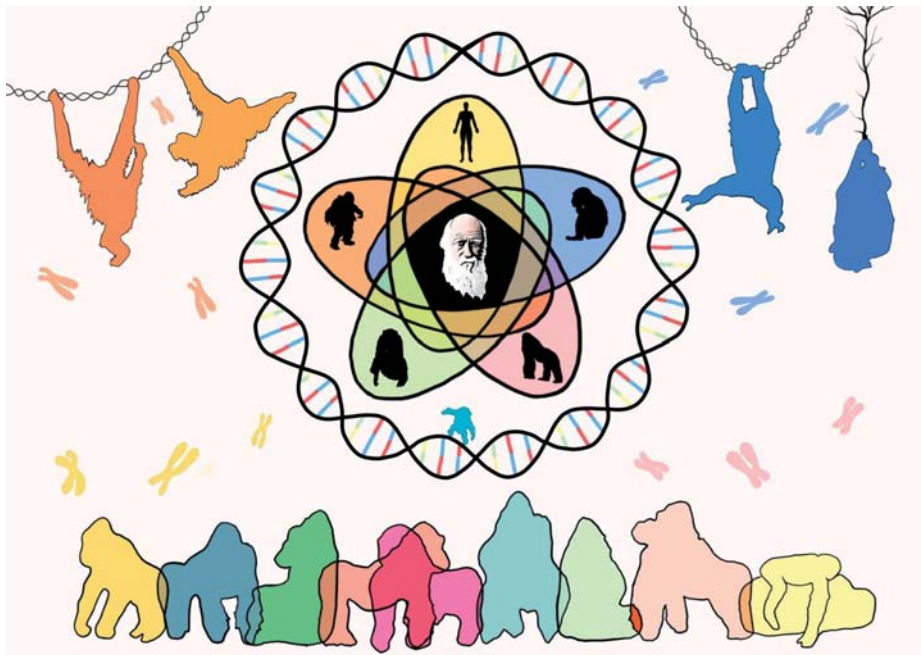
Chapter 3

Natural selection in the Great Apes

Published in Molecular Biology & Evolution (2016).

by

Alex Cagan, Christopher Theunert , Hafid Laayouni H, Gabriel Santpere, Marc Pybus, Ferran Casals, Kay Prüfer, Arcadi Navarro, Tomas Marques-Bonet, Jaume Bertranpetit, Aida M Andrés



Natural Selection in the Great Apes

Alexander Cagan,^{†,1} Christoph Theunert,^{†,1,2} Hafid Laayouni,^{†,3,4} Gabriel Santpere,^{†,3,5} Marc Pybus,³ Ferran Casals,⁶ Kay Prüfer,¹ Arcadi Navarro,^{3,7} Tomas Marques-Bonet,^{3,7} Jaume Bertranpetit,^{†,3,8} and Aida M. Andrés^{*,†,1}

¹Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

²Department of Integrative Biology, University of California, Berkeley, Berkeley, CA

³Departament de Ciències Experimentals i de la Salut, Institut de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

⁴Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Catalonia, Spain

⁵Department of Neuroscience, Yale University School of Medicine, New Haven, CT

⁶Genomics Core Facility, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

⁷Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

⁸Department of Archaeology and Anthropology, Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Cambridge, United Kingdom

[†]These authors contributed equally to this work.

[‡]These authors equally co-supervised this work.

*Corresponding author: E-mail: aida_andres@eva.mpg.de.

Associate editor: Ryan Hernandez

Abstract

Natural selection is crucial for the adaptation of populations to their environments. Here, we present the first global study of natural selection in the *Hominidae* (humans and great apes) based on genome-wide information from population samples representing all extant species (including most subspecies). Combining several neutrality tests we create a multi-species map of signatures of natural selection covering all major types of natural selection. We find that the estimated efficiency of both purifying and positive selection varies between species and is significantly correlated with their long-term effective population size. Thus, even the modest differences in population size among the closely related *Hominidae* lineages have resulted in differences in their ability to remove deleterious alleles and to adapt to changing environments. Most signatures of balancing and positive selection are species-specific, with signatures of balancing selection more often being shared among species. We also identify loci with evidence of positive selection across several lineages. Notably, we detect signatures of positive selection in several genes related to brain function, anatomy, diet and immune processes. Our results contribute to a better understanding of human evolution by putting the evidence of natural selection in humans within its larger evolutionary context. The global map of natural selection in our closest living relatives is available as an interactive browser at <http://tinyurl.com/nf8qmzh>.

Key words: evolution, adaptation, comparative genomics, primates.

Introduction

Understanding the adaptive genetic changes that led to the emergence of modern humans continues to be a major focus of modern genomics (Pritchard et al. 2010; Enard et al. 2014). However, despite much work in this field, many central questions remain unanswered. For example, it is still unclear what percentage of the human genome has been shaped by natural selection, which genetic variants are responsible for the phenotypes that make humans unique, and to what extent demographic factors have influenced the rate of adaptive evolution through human history. These questions can only be answered through a deeper understanding of the evolution both of the human genome and also of other closely related species. While laboratory studies on adaptation in organisms such as *Drosophila* have furthered our understanding of adaptive evolution (Lee et al. 2014), the usefulness of

these model organisms for understanding adaptation in humans is limited by the wide disparities that exist between them and humans, in both physiology and demography. Investigation of the molecular basis of adaptation is also hindered by differences in the structure and content of the genomes of more distantly related organisms. Studying our closest living relatives, the great apes, is therefore crucial for furthering our understanding of human evolution.

The *Hominidae* (humans and great apes) share several traits that make them particularly interesting. Relative to their ancestors they have evolved larger brains, more complex social systems and, arguably, the ability to create and maintain cultural traditions (McGrew 2004). Furthermore, the *Hominidae* species differ from one another in important ways (including their morphology, physiology, behavior and

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

life history traits) that may result from their independent adaptation to particular environments. Evolutionary genomic information can help us to understand the origin and molecular bases of both shared and species-specific traits in the *Hominidae*.

The *Hominidae* also provide an excellent system for comparative studies. This is because although the species are very closely related (with all lineages diverging over the last 12 My) they differ substantially in relevant features such as the effective size of their populations (N_e) (Prado-Martinez et al. 2013). This makes them well-suited for addressing longstanding theoretical questions in evolutionary biology. A central principle of population genetics is that the effective size of a population influences the efficacy of selection (Charlesworth 2009). Populations with a large N_e are expected to be more efficient at both fixing beneficial alleles and removing deleterious ones, when compared with populations with small long-term N_e or that have experienced severe bottlenecks. Empirical attempts to quantify this effect have been limited, with exceptions that include work in yeast (Elyashiv et al. 2010), *Drosophila* (Jensen and Bachtrog 2011) and eukaryotes (Grossman et al. 2013), as well as comparisons of very divergent lineages (Corbett-Detig et al. 2015). It remains unclear to what extent differences in N_e between closely related mammalian species impact the process of natural selection (Ellegren and Galtier 2016). Full genome sequences of humans and great apes provide a unique opportunity to investigate this question over a relatively short evolutionary timescale.

The signatures of natural selection have been extensively studied in humans (Bustamante et al. 2005; Sabeti et al. 2006; Nielsen et al. 2009; Andrés et al. 2010) and some of the apes (Mikkelsen et al. 2005; Locke et al. 2011; Prüfer et al. 2012; Sully et al. 2012; Bataillon et al. 2015; McManus et al. 2015). However, no study has comprehensively investigated the evidence for natural selection across the *Hominidae* lineages. We analyzed whole-genome sequence data from multiple individuals from lineages covering all major *Hominidae* species and subspecies (except *Gorilla beringei beringei*) (Prado-Martinez et al. 2013) and present the first investigation of the impact of natural selection using this dataset. We focus on attributes of the data that allow us to detect the different types of selection across evolutionary timescales. We then integrate these results to investigate the influence of N_e on the efficacy of natural selection, the targeted functional elements, the genes and biological processes targeted by each type of selection, and the conservation of selective pressures across *Hominidae* lineages.

Results

Sample Processing

In order to assess the influence of natural selection, we use a dataset of 54 non-human great ape and nine human genomes sequenced to an average of 25-fold coverage (supplementary table S1, Supplementary Material online). Because of differences in demography and selective pressures on autosomes and sex chromosomes, we focus exclusively on the autosomes.

We take particular care to minimize the influence of errors and biases in genomic data and ensure that our data is of the highest possible quality—something particularly important when comparing species. All reads were mapped to the same reference genome (human hg18). We built on the extensive data filtering strategy of Prado-Martinez et al. (2013) (see “Dataset” in “Methods” section). This conservative filtering strategy resulted in the exclusion of 726 Mb (~23%) of the autosomal genome. This includes tandem repeats (~38 Mb), segmental duplications (~154 Mb) and structural variants annotated in at least one species (~334 Mb) (see supplementary fig. S1, Supplementary Material online), all identified by unusual read-depth, so alternative methods (Gokcumen et al. 2013; Sudmant et al. 2015) may identify nonidentical regions. While certain genomic regions and gene families may be enriched in structural variation and be disproportionately affected by this filtering step, their removal is essential to minimize artifacts. We also excluded genomic gaps (~226 Mb) and base pairs that were not covered by a minimum of five reads in all individuals per species. The resulting dataset includes on an average 2,099 Mb of analyzable genome sequence per species (see supplementary fig. S1, Supplementary Material online). Although every filtering strategy has limitations and putative biases, we aim for a conservative approach that minimizes the presence of artifacts. The result is a high-quality comparative genomic dataset that allows us to investigate the signatures of natural selection and compare them across species (see supplementary materials Sample Processing, Supplementary Material online).

Neutrality Tests

We selected a set of neutrality tests that explore different aspects of the patterns of polymorphism and in combination allow us to detect the signatures of different types of natural selection across different time depths, from the emergence of the *Hominidae* ~12 Ma to recent and ongoing species-specific selective sweeps (fig. 1). Many neutrality tests exist; among them we chose those that utilize the type of information that we have (i.e., that do not require phased genomes), that have been shown to have high power to detect selection (Zhai et al. 2009), that explore relatively independent signatures and that provide information on different timescales. To keep the analyses manageable, we focus on four tests (see fig. 2):

- To detect signals of purifying and positive selection on the coding sequences of proteins, we applied the McDonald–Kreitman test (MK test; McDonald and Kreitman 1991). The MK test is run on a protein-coding gene-by-gene basis (supplementary materials MK 1, Supplementary Material online). By using information on sequence divergence, it has power to detect signatures of positive and purifying selection along the entire branch lengths of the *Hominidae*.
- To detect long-term balancing selection and positive selection that could have occurred at a deep evolutionary time, we applied a statistic based on the Hudson–Kreitman–Aguadé test (HKA; Hudson et al. 1987), which has been found to be a highly powerful method to detect

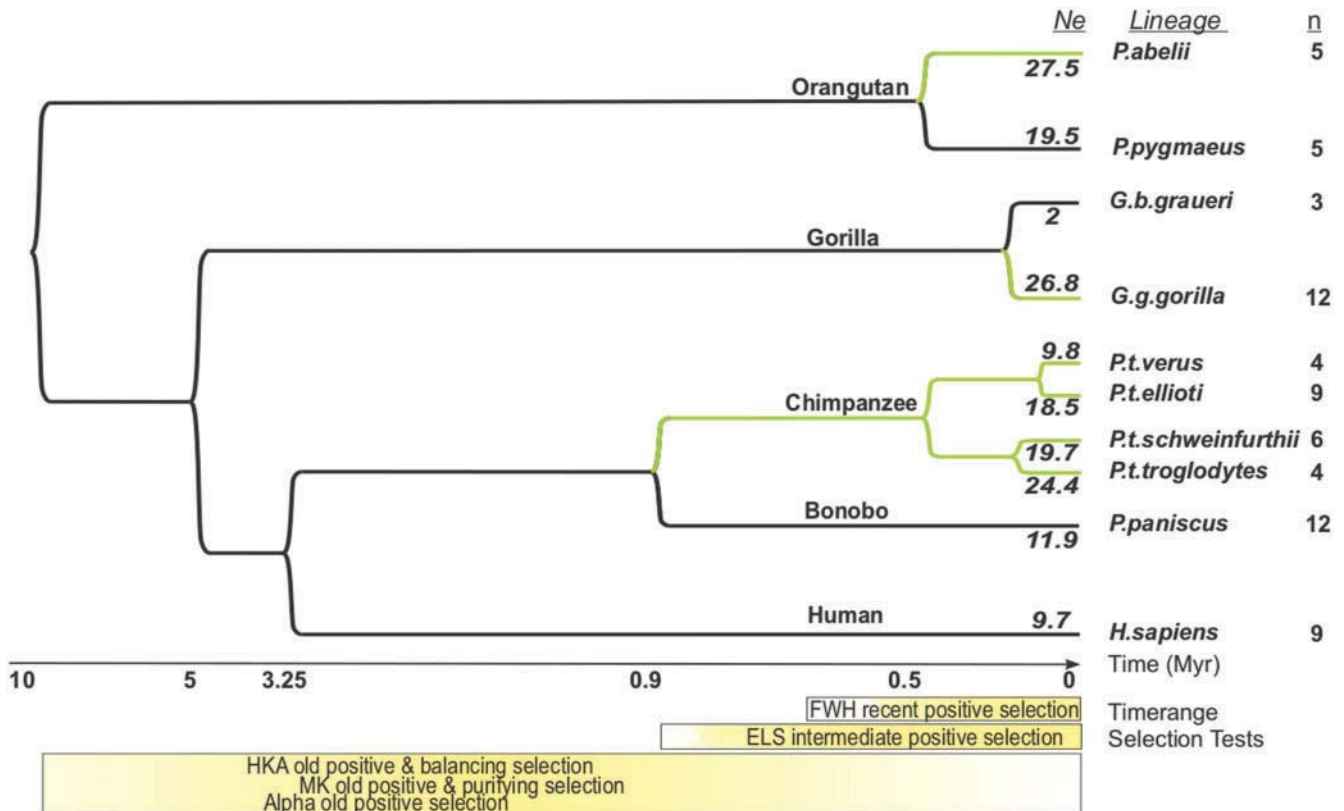


Fig. 1. Timescale of neutrality tests. *Hominidae* phylogeny with the approximate time ranges where each neutrality test has power to detect signatures of natural selection. (a) The lineages with the number of genomes used in this study are shown on the right. The X-axis shows the timescale, in units of millions of years. Split times of lineages from Prado-Martinez et al. (2013). For FWH, MK and HKA, the approximate time range where the tests are inferred to have most power to detect selection are represented by color intensity. For ELS, we label in green the branches where the test has power to detect selection. n: number of individuals in each lineage. Ne: estimates of effective population size in units of thousands of individuals according to Watterson's estimator, taken from Prado-Martinez et al. (2013).

positive selection (Zhai et al. 2009). The HKA statistic was calculated across the genome in 30-kb windows with a 15-kb overlap between windows (Methods and supplementary materials HKA 1, Supplementary Material online). As it uses both divergence and diversity data, the HKA statistic has power to detect positive selection over broad timescales as well as long-term balancing selection, including persistent balancing selection that predates the emergence of the *Hominidae*, such as on the MHC region (Hedrick 1999).

- To detect lineage-specific positive selection that occurred after the divergence of an ancestral population into two species, we applied the Extended Lineage Sorting test (ELS; SOM 13 in Green et al. 2010; Supplementary Information 7 in Prüfer et al. 2012; Supplementary Information 19a in Prüfer et al. 2014). The test is run across the genome and identifies regions without a pre-defined size (Methods and supplementary materials ELS 1, Supplementary Material online).
- To detect recent selective sweeps, we applied Fay and Wu's H statistic (FWH; Fay and Wu 2000). The FWH statistic was calculated across the genome in 30-kb windows with 15-kb overlap between windows (Methods and supplementary materials FWH 1, Supplementary Material online).

Together, these tests detect the signatures of purifying, balancing and positive selection, old and recent (we refer to events in the order of millions of years for old and of hundreds of thousands of years up to present day for recent selection), in each lineage (fig. 1). Integrating all results provides an unprecedentedly broad picture of the targets of natural selection in the genomes of humans and great apes.

Ne and the Strength of Natural Selection in the Great Apes

As discussed earlier, empirical data is limited regarding the effect of long-term Ne on the efficacy of natural selection over short evolutionary timescales in vertebrates. The relationship between population size, selection and levels of neutral diversity in populations continues to be a matter of considerable debate (Ellegren and Galtier 2016). A recent study (Corbett-Detig et al. 2015) proposed that the effects of linked selection can explain Lewontin's paradox (1974), namely that neutral diversity does not scale as expected with population size. Though a recent reanalysis of this data suggests that while linked selection influences diversity along genomes, fluctuations in Ne are the major driver of levels of diversity between species (Coop 2016). The debate has so far been hampered by the limited availability of population-level genome sequence

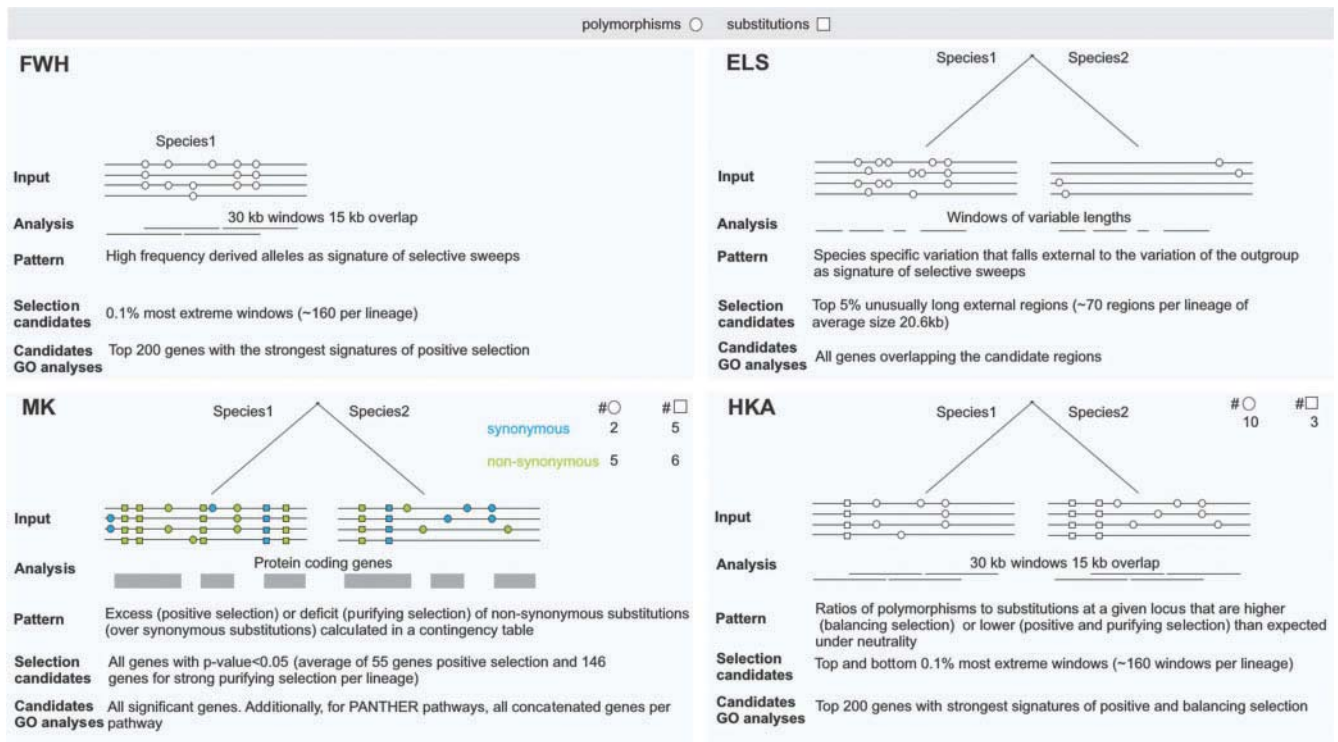


Fig. 2. Summary of the neutrality tests used. Each box presents the input (the information used), the analysis strategy (how each test was applied on the genome-wide data), the pattern (the signatures of selection explored), the criteria to select selection candidates (the top candidates for each test) and the criteria to select candidates for GO analyses (the candidate used for gene ontology analyses).

data across species (Ellegren and Galtier 2016). Our dataset therefore provides an ideal opportunity to investigate the relationship between N_e and selection in closely related species.

We find that the ratio of nonsynonymous to synonymous substitutions negatively correlates with long-term N_e in this dataset (Prado-Martinez et al. 2013), as expected with more efficient purifying selection in populations with higher N_e . Here we aim to: (1) infer the distribution of fitness effects (DFE) in each species, (2) quantify the magnitude of the influence of N_e on the DFE, (3) compare the influence of long-term versus short-term N_e , and (4) investigate its influence not only on purifying, but also on positive selection.

Ne and the Strength of Purifying Selection

We first inferred the DFE of deleterious mutations for 3,859 one-to-one orthologous protein-coding genes with DFE-alpha (Eyre-Walker and Keightley 2009), which is based on the MK test (see Methods and MK supplementary materials, Supplementary Material online). The method fits a demographic model to the SFS of neutral sites, and, simultaneously, estimates the gamma-distributed DFE of new nonneutral mutations and the fraction of adaptive substitutions (α) (supplementary fig. S16, Supplementary Material online). For all lineages, the shape parameter of the gamma distribution is < 1 , indicative of highly leptokurtic (L shaped) DFEs and most nonsynonymous mutations being strongly deleterious. Indeed, in all lineages the proportion of nonsynonymous mutations with a $N_e S > 10$ (S being the mean homozygous effect

of a deleterious variant) is $> 65\%$ (supplementary table S104 and fig. S18, Supplementary Material online), similar to estimates for humans (Eyre-Walker and Keightley 2009) and gorillas (McManus et al. 2015). We observe that the proportion of predicted neutral or nearly neutral mutations correlates negatively with long-term N_e (correlation of -0.64 , P value = 0.04 after accounting for phylogenetic nonindependence using BayesTraitsV2 random walk/maximum likelihood method; Pagel and Meade 2013). This correlation reflects stronger purifying selection in great ape species with a larger long-term N_e .

Efficient purifying selection reduces also the accumulation of linked genetic variation due to background selection. Within the bins in the middle range of the HKA distribution (see “Methods” section), which are particularly sensitive to purifying selection, lower HKA scores associate with stronger background selection (lower B scores, supplementary fig. S6, Supplementary Material online) and a higher proportion of protein-coding exons (fig. 3). This is expected if background selection reduces diversity around protein-coding and other functional regions. Across lineages, and in agreement with the DFE-alpha results, the effects of purifying selection increase with larger N_e both when considering the proportion of protein-coding exons and the B scores (supplementary materials HKA 3 and supplementary table S5, Supplementary Material online) (McVicker et al. 2009). Incidentally, the effect is much weaker for nonprotein coding exons (supplementary table S5 and supplementary fig. 1E, Supplementary Material online). These results are virtually unchanged if we use only lineages with less than ten individuals or only lineages with

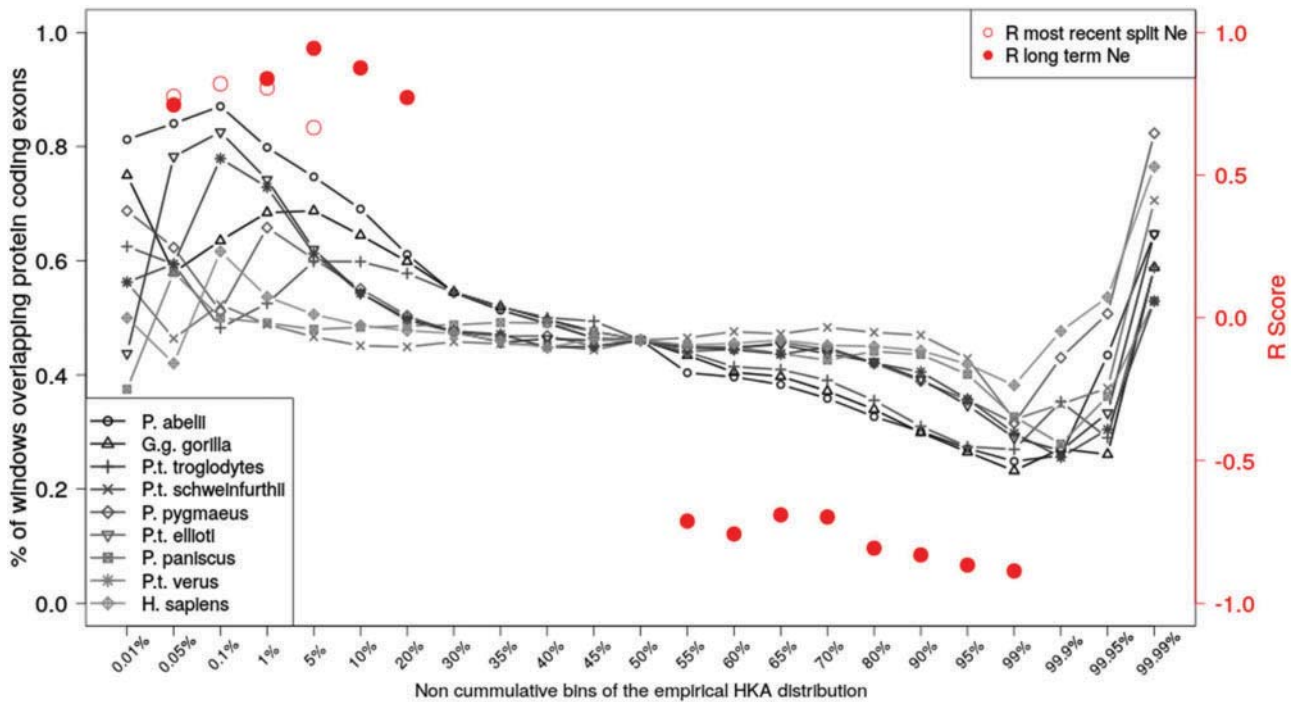


Fig. 3. Percentage of windows overlapping protein coding exons. Percentage of windows overlapping protein coding exons for noncumulative bins of the HKA empirical distribution (X-axis). Each lineage is plotted as a shaded line. The Pearson's correlation (R) between the percentage of windows overlapping protein coding exons and N_e within each HKA bin and across all lineages is shown on the right Y-axis. Pearson's correlation coefficient was computed both with an estimate of short- and long-term N_e (from Prado-Martinez et al. 2013). Only R values with significant P values ($P < 0.05$) are shown.

more than five individuals, suggesting that sample size differences between lineages do not affect our observations (supplementary materials subsampling analysis 1.4 and supplementary table S106 and supplementary fig. S30, Supplementary Material online).

The correlations with N_e above are almost always stronger with long-term N_e than with recent N_e (for 21 of the 23 HKA bins; supplementary table S3, Supplementary Material online), indicating that we detect the effects of long-term evolutionary history rather than only differences in power due to overall levels of diversity (although differences in the accuracy of the N_e estimates may affect this comparison). Therefore, despite the recent and ongoing population declines experienced by many of these species, their long-term N_e appears to be a better predictor than recent N_e of the past efficacy of purifying selection.

Ne and Adaptive Evolution

With the MK-based DFE-alpha, it is possible to estimate the proportion (α) of nonsynonymous substitutions driven by positive selection, as well as the ratio of adaptive to neutral divergence ($\omega(\alpha)$) (supplementary table S103, Supplementary Material online). With the exceptions of *Pongo pygmaeus* (with poor bootstrap support), and *Pan t. schweinfurthii* (where two inbred individuals (Prado-Martinez et al. 2013) dramatically increase the estimates) (supplementary table S103, Supplementary Material online), both the estimated proportion (α) and the estimated rate of adaptive evolution are low (0–12% and 0–2%, respectively)

in agreement with previous estimates (Eyre-Walker 2006; McManus et al. 2015). We observe that in nonhuman great apes both the proportion and the rate of adaptive substitutions are positively correlated with long-term N_e , after phylogenetic nonindependence is accounted for using a generalized least square approach (supplementary fig. S17 and supplementary materials MK test 2.3, Supplementary Material online). The correlation is high and significant when all nonhuman species, except the problematic *Pongo pygmaeus* and *Pan troglodytes schweinfurthii*, are considered (Pearson's $R = 0.9$, P value = 0.004) (see fig. 4).

The effect of positive selection on linked variation also increases with long-term N_e . In the bottom bins of the HKA empirical distribution, which are enriched in targets of positive selection, the percentage of protein-coding windows correlates positively with N_e (0.05–1% bins, P values = 0.0001–0.02). Only the lowest 0.01% HKA bin is not significant, potentially due to the spatial clustering of windows as a result of selective sweeps (supplementary materials HKA 4 and supplementary table S12, Supplementary Material online).

Thus, our results indicate that long-term N_e of populations significantly affects the efficacy of both purifying and positive selection. These correlations are remarkable because these species are very closely related and their long-term N_e varies by a maximum difference of 3-fold.

The Candidate Targets of Natural Selection

As most genomic sites evolve neutrally or nearly neutrally (Kimura 1979; Kelley et al. 2006), we expect an enrichment

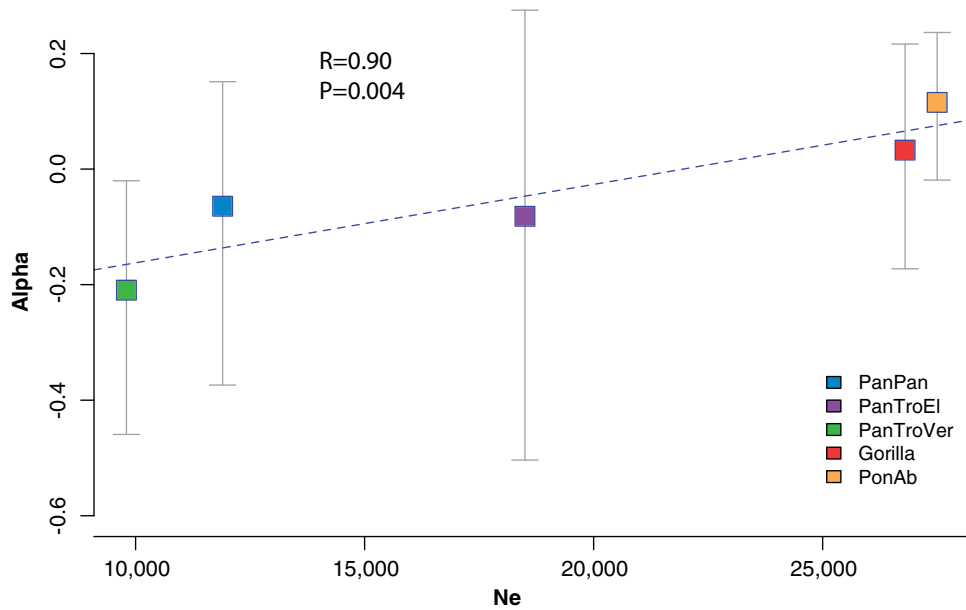


Fig. 4. Correlation between rate of adaptive substitutions (α) and effective population size (N_e). The X-axis shows the effective population size. On the Y-axis, the rate of adaptive substitutions is plotted as α . Correlations were calculated while controlling for the phylogenetic nonindependence using a generalized least square approach and a random walk/maximum likelihood method (see [Supplementary Materials MK 2.3](#), [Supplementary Material](#) online).

of targets of natural selection in the extreme tails of the genome-wide distributions of neutrality test statistics. Therefore, we can identify candidate targets of natural selection without relying on a simulated neutral expectation, which is vulnerable to parameter misspecification, an important problem given the complex evolutionary history of the *Hominidae* lineages. Given the little we know about the strongest targets of purifying, positive and balancing selection in nonhuman apes, this catalog is highly relevant. In addition, these loci allow us to investigate the tempo, conservation, and biological function of natural selection in the great apes. The nature of each of the tests considered means that their implementation in the genome varies and the criteria to define candidate targets of selection necessarily varies too (see [fig. 2](#)).

Sample Size and the Candidate Targets of Natural Selection
Sample size, which varies among lineages, may influence the power to detect signatures of natural selection. We assess how differences in sample size might influence our results with down-sampling analyses. We randomly down-sample, 100 times, four or eight individuals from the two lineages with the largest sample size (*Pan paniscus* and *Gorilla gorilla gorilla*) and run all neutrality tests for chromosome 1 (except the ELS test for *Pan paniscus* because this test is not appropriate for this lineage). We then measure the overlap between the candidates from the down-samples (0.1% or 1% tail of the empirical distribution) with the equivalent candidates from the original results ([supplementary materials](#) subsampling analysis 1, [Supplementary Material](#) online).

The impact of sample size differs between selection tests. HKA appears very robust to sample size variation for signatures of positive selection, showing a mean overlap between the original and down-sampled results of at least 86%

([supplementary materials](#) subsampling analysis 1.1 and [supplementary figs. S21–24](#), [Supplementary Material](#) online). This is likely to be because the HKA is not strongly affected by the allele frequency of polymorphisms.

In contrast, FWH and ELS appear more sensitive to sample size ([supplementary materials](#) subsampling analysis 1.2–1.3 and [supplementary figs. S25–28](#), [Supplementary Material](#) online). This may be due to the influence of sample size on the estimates of allele frequency. Therefore, we find that the HKA test is better suited for comparative analyses where sample sizes are low or unequal between populations.

An Available Genome-Wide Map of Natural Selection in Hominidae

The genome-wide map of signatures of natural selection includes information about different types of selection over varying time frames. As such, it provides a broad picture of the influence of natural selection in each genomic region and *Hominidae* species. All the information is available as an interactive browser at webpage: <http://tinyurl.com/nf8qmzh> following the criteria and configuration of a recently published human dataset (Pybus et al. 2014, 2015). The UCSC-style format facilitates the integration with the rich UCSC browser tracks, a search mask allows easy access to results for specific genes or genomic regions, and the raw scores (test statistic value and rank score/empirical *P* value) can be conveniently downloaded using the UCSC Table function. We expect this to be a valuable resource for a wide range of analyses.

The Functional Targets of Natural Selection

The relative contributions of variants in regulatory versus protein coding regions of the genome to adaptive evolution

remains a matter of controversy (Halligan et al. 2013). Since King and Wilson (1975) the relative importance of coding and regulatory variation to adaptive evolution has been contentious. Protein-coding DNA constitutes ~1.5% of the genome but 10–15% appears to be functionally constrained (Ponting and Hardison 2011). The role of nonprotein-coding genes and other nongenic elements in genome function and evolution remains debated (Encode Project Consortium 2011; Doolittle 2013) with several lines of work suggesting that nongenic regions (including some gene deserts) can play an important role in phenotype and adaptation (Bejerano et al. 2006; Libioulle et al. 2007; McPherson et al. 2007; Hubisz and Pollard 2014). Although the stringent filtering of the data and the imperfect annotation of nonprotein-coding functional elements hampers the comparison of protein-coding versus nonprotein regions, we investigated the functional annotations of our candidates.

Except for MK, the neutrality tests we used are agnostic about functional annotation. Still, most of our candidate targets of positive selection contain functional annotations: mean values across species are 72% for HKA, 71% for ELS and 80% for FWH. This is significantly greater than genome-wide expectations based on random sampling of the callable genome (P values < 0.05 in all lineages except *Pan paniscus*, P value = 0.2, and *Pongo pygmaeus*, P value = 0.11) (supplementary materials HKA 7 and supplementary figs. S7–15, Supplementary Material online). Among these annotations, the overlap with protein-coding exons (HKA = 62%, ELS = 59%, FWH = 45%) is significantly enriched in all lineages except *Pan paniscus* (P value = 0.15) and *Pan troglodytes verus* (P value = 0.06). In contrast, the mean overlap with exons from nonprotein coding genes (HKA = 18%, ELS = 2%, FWH = 20%), is not significantly elevated relative to genome-wide levels (supplementary figs. S7–15, Supplementary Material online, lowest P value = 0.16 in *Gorilla gorilla gorilla*).

Candidate targets of balancing selection are also highly enriched in protein-coding exons (e.g., 64% in the top 0.01% bin) and in the top HKA bins this proportion sharply increases with the HKA score (fig. 3). The increased levels of diversity in these windows cannot be explained by technical artifacts, as these regions are not unusual in terms of coverage or mapping quality (supplementary materials HKA 1.3–1.4, Supplementary Material online) or by current models of neutral evolution or purifying selection, and are instead best explained by long-term balancing selection acting on or near these protein-coding exons.

The Biological Pathways Targeted by Natural Selection

According to our results above, protein-coding genes appear to be a key target of natural selection in the *Hominidae*. We thus investigated the biological functions that these genes are involved in. For each neutrality test and lineage, we identified the genes in candidate regions of positive or balancing selection (see fig. 2 and “Methods” section for details) and performed gene enrichment analyses using WebGestalt (Zhang et al. 2005). Our necessarily strict data filtering may

disproportionally affect certain Gene Ontology (GO) categories (e.g., olfactory receptors), but we discuss below the categories that retain the strongest signatures for each type of natural selection.

Pathways Targeted by Balancing Selection

The top genes for balancing selection include a number of well-established targets, such as the major histocompatibility complex (MHC) genes (Hughes and Yeager 1998; Hedrick 1999). Indeed, windows containing MHC genes appear among those with the strongest signals of balancing selection in all lineages (supplementary tables S6 and S13, Supplementary Material online). In addition, in all lineages there is a significant enrichment of immunity-related categories such as the GO “Antigen processing and presentation” category (closely related to the MHC) (supplementary tables S16–40, Supplementary Material online). This provides evidence that balancing selection has a strong influence on immunological pathways in all lineages.

To test whether there were strong signatures of balancing selection beyond the MHC complex, we re-ran the GO enrichment analysis excluding all genes in the MHC region on chromosome 6 (supplementary tables S57–65, Supplementary Material online). Doing so removes the enrichment for the GO category “Antigen processing and presentation” in all lineages. Interestingly, in three of the four *Pan troglodytes* lineages (excluding *Pan troglodytes schweinfurthii*) there is significant enrichment for the GO category “Cornified envelope” (supplementary tables S360, S62, and S63, Supplementary Material online), driven by the three genes *SCEL*, *SPRR2B* and *SPRR2G*. The cornified envelope is the most exterior layer of the skin and consists of dead cells. Related to this, we note that in *Pan troglodytes verus* the GO category “keratinocyte differentiation”, involved in the development of the most common cell type in the epidermis is also significantly enriched (P value = 0.0026). This is interesting because keratins and proteins similarly involved in epithelial barrier formation have been proposed as targets of balancing selection (see “Discussion” section).

Pathways Targeted by Strong Purifying Selection

Since *Hominidae* are closely related, we would expect that similar regions are evolving under purifying selection. We therefore tested whether the pathways showing the strongest signatures of constraint are consistent among lineages. From the MK test, 53 of the 152 evaluated pathways showed signatures of strong purifying selection in more than one lineage. In particular, the “Integrin signaling” pathway and “Wnt signaling” pathway, which regulate basic cellular and developmental processes and the “Alzheimer’s disease-presenilin” pathway are significantly constrained across all lineages (supplementary table S102, Supplementary Material online).

Pathways Targeted by Positive Selection

For the HKA, several lineages show evidence of positive selection targets being enriched for GO categories related to immune function. For example, the GO category

“Complement activation” (genes that activate the innate immune system) is significantly enriched in *Pan paniscus* (P value = 0.042; all P values adjusted for multiple testing), whereas the related pathway “Complement receptor activity” is enriched in *Pongo abelii* (P value = 0.011) and the GO category “Viral receptor activity” in *Gorilla gorilla gorilla* (P value = 0.0004) (supplementary tables S41, S46, and S54, Supplementary Material online).

We find that the FWH candidate targets of positive selection show enrichment in several GO categories related to brain development and function, exclusively within the African *Hominidae* lineages. This includes for example the GO categories “Dendrite” (*Homo sapiens* P value = 0.040; *Pan troglodytes troglodytes* P value = 0.010; *Gorilla gorilla* P value = 0.0024) and “Neuron spine” (*Pan troglodytes troglodytes* P value = 0.001; *Gorilla gorilla gorilla* P value = 0.006). Several additional neurological categories are enriched in single lineages (supplementary tables S75–86, Supplementary Material online). For example, *Homo sapiens* is the only lineage with significant enrichment of the GO category “Glutamate receptor activity” (P value = 0.002); glutamate is the main excitatory neurotransmitter in the brain.

For the MK, the set of genes with an excess of divergence is small (supplementary table S96, Supplementary Material online) but we found a significant enrichment in genes involved in, for instance, “Ion channel activity” in *Pan paniscus* (P value = 0.034), and in “Glycosaminoglycan biosynthesis” in *Gorilla gorilla gorilla* (P value = 0.020), among other pathways (supplementary tables S98–101, Supplementary Material online).

Overlap between Targets of Positive and Balancing Selection across Lineages

The *Hominidae* lineages have shared, along their evolutionary history, similar physiologies and environments. As such, they have likely been subject to common selective pressures even after their lineages split. To investigate this possibility, we identified genes that show similar signals of natural selection in multiple lineages. Since we use an empirical approach to identify candidate targets of natural selection (as the demographic models for these species are not well established), we use the same 0.1% cut-off to identify outliers from both tails of the HKA empirical distribution and one tail of the FWH distribution. Therefore, we cannot make general claims about the relative frequency of positive and balancing selection in primate species. We can though explore the level of sharing across lineages of these candidate targets. To be conservative, we only consider genes to be shared targets of selection if they appear as candidates in at least three lineages.

Overlap between Selection Targets

We find no signals of positive selection that are shared across all lineages. In fact, there is modest sharing across lineages, a possible indication of the lineage-specific nature of the adaptive process (although we note that we are highly conservative in our selection of candidate genes and the power of FWH is reduced with lower sample sizes) (supplementary

table S15, Supplementary Material online). We observe that the HKA candidates show lower sharing across lineages than those from the FWH (supplementary tables S14 and S73, Supplementary Material online). For the HKA, only 27 genes (of the 200 candidates per lineage) are shared in at least three lineages compared with 67 for the FWH, which detects more recent selective events. We note that shared signals among the *Pan troglodytes* sub-species may not reflect truly independent signatures of selection as signals may predate their divergence into separate lineages and the possibility of admixture between sub-species. However, of these 67 genes, only a minority (8) are shared exclusively among the *Pan troglodytes* subspecies, potentially reflecting their recently shared ancestry. The majority (59) show evidence of recent positive selection across a range of lineages (supplementary table S73, Supplementary Material online) suggesting putative parallel adaptive events.

Turning to the biological function of these genes, we find limited enrichment among HKA targets (the only significant GO category is “Structural molecule activity”, P value = 0.009; supplementary table S3AAB, Supplementary Material online). However, the 67 shared FWH candidate genes are significantly enriched in multiple functional categories (supplementary tables S87–89, Supplementary Material online), including several neuronal pathways, suggesting that these are a common target of recent positive selection across the *Hominidae*.

Genes targeted by balancing selection show much greater sharing across lineages, with 156 genes showing signatures of balancing selection across at least three lineages (supplementary table S13, Supplementary Material online, fig. 5 for an example across *Pan troglodytes* lineages). Nine genes, primarily from the MHC region, are shared across all lineages (supplementary table S13, Supplementary Material online). This likely reflects the long-term and persistent nature of balancing selection on immunity-related genes, in particular in the MHC.

Discussion

We present a global investigation of the signatures of purifying, positive, and balancing selection at different time scales and across the *Hominidae* lineages. We observe strong evidence for the signatures of each of these types of natural selection on patterns of genomic variation. By carefully avoiding technical differences, we can compare, for the first time, the patterns of different types of natural selection across the great ape species. All genomic analyses of signatures of selection are complicated to some extent by demographic processes which can result in patterns of genomic variation that obscure signatures of selection or produce false-positives. We tried to mitigate this by utilizing tests that identified putatively selected regions as outliers based on the entire distribution of patterns of genomic variation, under the assumption that the majority of the genome is evolving neutrally.

We find that even with the relatively similar N_e of the great apes (with a maximum difference of 3-fold), N_e has a significant effect on the efficacy of natural selection. This appears to be true for both purifying and positive selection. The

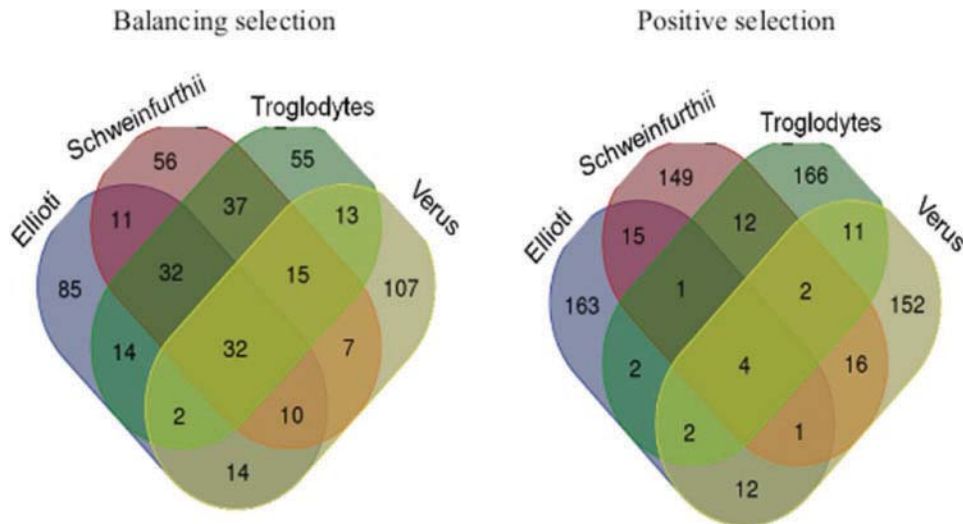


Fig. 5. Venn Diagram of shared targets of balancing and positive selection among *Pan troglodytes* lineages. Overlap of the number of putative target genes of balancing and positive selection as inferred by the HKA test for all *P. troglodytes* lineages.

evidence for adaptive evolution is stronger in protein-coding than in nonprotein coding genes, and it is overrepresented not only on loci with relevance to immune function, but also on loci involved in the development and maintenance of the brain. Long-term balancing selection, which most clearly affects the evolution of immune and skin-related loci, is more often shared across lineages than positive selection. In what follows, we briefly discuss these observations, as well as some of the biological insights from the loci identified.

Effective Population Size Significantly Influences the Efficacy of Natural Selection in the *Hominidae*

We estimate that at least 65% of mutations are deleterious ($NeS > 10$) in all lineages. This estimate agrees very well with results in humans (Eyre-Walker and Keightley 2009). Our results also agree with a recent study in gorillas (McManus et al. 2015) where the DFE-alpha method provided a very similar gamma shape parameter and proportion of strongly deleterious and neutral alleles (66% and 23.8%, compared with our 65% and 22%). Regarding the prevalence of positive selection, our estimates in *Gorilla gorilla gorilla* (3%) also overlap with previous estimates (McManus et al. 2015). Our results thus confirm the limited information that exists for the *Hominidae* and greatly expand upon it.

Having information for many great apes enables us to start to compare the different species. The strength of purifying selection on nonsynonymous sites, and its effects on linked variation, correlate with the long-term Ne of the populations. Similar correlations have been observed among other, more distant, species (Leffler et al. 2012; Corbett-Detig et al. 2015). However, our results indicate that even the modest Ne differences that exist among the *Hominidae* have also affected the efficacy of positive selection, which is likely to be less prevalent and more dependent on environmental changes than purifying selection. Therefore, the *Hominidae* lineages with the largest long-term effective population sizes, such as *Gorilla gorilla gorilla* and *Pan troglodytes troglodytes*, are

better able to both remove deleterious alleles and fix adaptive alleles than lineages such as *Pan paniscus* and *Homo sapiens*. Even the relatively recent differences in long-term Ne between *Pan troglodytes* sub-species seem to have resulted in differences in the efficacy of natural selection. This may be extremely important because the long-term survival of these species, which live in small populations and are endangered, may depend on their ability to adapt to changes in their local environments.

Biological Interpretation of Candidate Genes

Our data indicates that adaptive evolution (positive and balancing selection) often targets protein-coding regions (be that the protein-coding sequence or the surrounding regulatory elements). This suggests that in these species variants that affect proteins have been important drivers of novel adaptations. It also supports the comparison of protein-coding versus nonprotein coding regions to establish patterns of positive selection, as long as additional confounding factors are accounted for (Coop et al. 2009; Key et al. 2016).

The candidate genes targeted by positive selection show only moderate overlap between species. This is not surprising, as different populations likely adapt differently at the genetic level even to similar environmental pressures. Nevertheless, there are certain genes, and even gene categories, that show evidence of positive selection in several lineages, which could reflect recurrent evolution at the genomic level. We note that the sharing across lineages is substantially higher when we turn to balancing selection. This is expected because we target only long-term balancing selection, which may predate the divergence of the great ape lineages.

There are several possible interpretations of shared signals. When the signature is shared across closely related lineages these most likely reflect shared events. When the signature is shared across distant lineages, this may reflect independent adaptive evolution. In these cases, selection may be favoring independent phenotypes in each species, for example if it affects different, neighboring functional elements in each

species or, due to pleiotropy, the same functional element for a different phenotype in each species. Alternatively, these regions may represent cases of convergent evolution, where the same phenotype is selected for across species. For example, many genes involved in brain development have shared evidence for positive selection across different species. We speculate that there has been ongoing positive selection for neurological phenotypes across the great apes and that although this was likely to be highly polygenic, some of the same genes may have been involved across species.

The detection of specific genes that have been under adaptive evolution is of great interest, especially when dealing with lineages closely related to humans. Here we discuss some of the most interesting findings. For an extended discussion of putatively selected genes, see [supplementary materials section 7, Supplementary Material](#) online.

Immunity

Balancing Selection on the MHC

Host–pathogen co-evolution can result in strong selective pressures ([Anderson and May 1982](#)). In agreement with this we find, in all lineages, evidence of balancing selection maintaining adaptive diversity on immunity-related genes. As expected with long-term balancing selection, where the time to the most recent common ancestor may predate the species split, many cases are shared among closely related lineages. These results provide further evidence that advantageous diversity is extremely important for the immune system, as has been shown in humans ([Ferrer-Admetlla et al. 2008](#); reviewed in [Key et al. 2014](#)). Not surprisingly, the MHC genes are among the top candidate targets of balancing selection in all lineages.

Balancing Selection on the Skin Barrier

The cornified envelope is a layer of dead keratinocytes (corneocytes) that are linked to structural proteins. They form a protective barrier in the outermost layer of the epidermis, known as the stratum corneum, which acts as an external wall that protects the body from physical injury and bacterial invasion. We find that the candidate targets of balancing selection are enriched in genes involved in keratinocyte differentiation in *Pan troglodytes verus*. They are also enriched in the related biological process “cornified envelope development” in *Pan t. ellioti*, *Pan troglodytes troglodytes* and *Pan troglodytes verus* (the genes involved are *SCEL*, *SPRR2B* and *SPRR2G*) and include two additional cornified envelope genes (late cornified envelope genes 3D and 3E, *LCE3D* and *LCE3E*). In *LCE3D* and *LCE3E*, the signatures are in flanking regions (the protein-coding portions of the genes are filtered out by the segmental duplication filter). We also identify *CDSN*, which encodes corneodesmosin, an adhesive protein involved in skin barrier integrity and has previously been shown to have signatures of balancing selection in humans ([Andrés et al. 2009](#); [Cagliani et al. 2011](#)), as a putative target of balancing selection in three lineages (*Homo sapiens*, *Pan paniscus*, *Pan troglodytes verus*) ([supplementary table S13, Supplementary Material](#) online).

A hypothesis for why genes involved in epidermal differentiation may evolve under balancing selection has been proposed in humans in relation to the filaggrin (*FLG*) gene ([Irvine and McLean 2006](#)), which is essential for the formation of the cornified envelope yet it has two common loss-of-function alleles (5% frequency each in Europeans) that cause ichthyosis vulgaris and strongly predispose to atopic dermatitis ([Irvine and McLean 2006](#); [Smith et al. 2006](#)). It has been proposed that the loss-of-function alleles might result in a leaky skin barrier through which low levels of pathogens can penetrate, promoting innate immunity through a process of “natural vaccination” ([Irvine and McLean 2006](#)).

Humans homozygous for loss-of-function *CDSN* alleles frequently show skin barrier defects and are susceptible to *Staphylococcus aureus* superinfections early in life, suggesting variation in the gene influences the ability of pathogens to penetrate the skin barrier ([Oji et al. 2010](#)). Interestingly, heterozygote carriers of this loss-of-function allele do not present these phenotypes, suggesting that heterozygotes may obtain benefits without the deleterious costs of homozygous carriers. Therefore, a leaky skin barrier that promotes “natural vaccination” may be a hitherto under-appreciated mechanism driving balancing selection on a variety of genes involved in development of the stratum corneum across species. We hypothesize that this mechanism may underlie the strong signatures of balancing selection we detect in *CDSN* (corneodesmosin) and other genes involved in the development of the cornified envelope. The presence of advantageous variation on genes involved in the formation of the epithelial barrier may therefore be more widespread than previously recognized.

Positive Selection on HIV/SIV-Related Genes

We also find evidence for pervasive positive selection on immune-related processes, as seen in *H. sapiens* and other *Hominidae* before ([Mikkelsen et al. 2005](#); [Cagliani et al. 2010](#); [Casals et al. 2011](#)). MK, HKA and FWH candidate targets of positive selection all are significantly enriched in genes related to immune response ([supplementary tables S16–55, S75–86 and S99, Supplementary Material](#) online). The particular genes vary between lineages, although some are shared across lineages.

Immunity-related genes with signals of selection in multiple lineages may reveal convergent adaptive response to pathogens, or adaptive introgression. The gene *IDO2* is identified as a FWH candidate of recent positive selection in all four *Pan troglodytes* lineages and *Pan paniscus*. *IDO2* encodes the enzyme indoleamine 2, 3-dioxygenase 2, which is involved in T-cell regulation and the Tryptophan oxidation pathway ([Metz et al. 2014](#)). This pathway is activated after HIV infection and causes chronic inflammation ([Murray 2010](#)), likely underlying HIV-1 immunopathogenesis ([Boasso and Shearer 2008](#)). Interestingly, blocking expression of the *IDO* genes in rhesus macaques infected with SIV/HIV improves health outcomes ([Boasso et al. 2009](#)). Therefore, selection on this functional pathway may contribute to the ability of some *Pan troglodytes* individuals to be resistant to AIDS progression after HIV infection, which has been attributed to a lack of

HIV induced T-cell dysfunction (Heeney et al. 1993). The MK test also identifies *HIVEP1* as a target of positive selection in *Pan troglodytes schweinfurthii* and *Pan paniscus*. The transcription factor encoded by *HIVEP1* binds enhancer elements of several promoters of viruses, including HIV-1. Investigation of these selection signals may be of relevance to treating HIV infections in humans.

In summary, we find strong evidence of shared signals of both balancing and, less frequently, positive selection on genes involved in immunity in the *Hominidae*. This likely reflects the strong and continuous selective pressure that infection and disease exerts on these populations, and the close evolutionary history of the *Hominidae*, which results in exposure to similar pathogens and similar genetic responses.

Neurological Functions

All *Hominidae* lineages are known to possess sophisticated cognitive abilities (Tomasello and Call 1997; McGrew 2004) related to their increased brain size and changes in brain organization relative to other primates (Semendeferi et al. 2002). We find some of the categories with the strongest evidence of purifying selection (with MK) are involved in brain function. It is intriguing that the candidate targets of recent positive selection are also enriched in neurological functional categories, with some genes involved in brain development and function showing signatures across multiple lineages.

The gene with signatures of positive selection across the highest number of species and timescales is *NRXN3*, which codes for neurexin 3. The gene is mainly expressed in the brain and encodes for a protein involved in synaptic transmission and plasticity; it belongs to a gene family associated with several cognitive diseases (Südhof 2008). *NRXN3* shows signatures of positive selection in six lineages, with a FWH signal of recent positive selection in *Pan troglodytes ellioti*, *Pan troglodytes schweinfurthii*, *Pan troglodytes troglodytes* and *Pongo pygmaeus*, an HKA signal of positive selection in *Homo sapiens*, and an ELS signature in all lineages where it was performed (*Pan troglodytes*, *Gorilla gorilla gorilla* and *Pongo abelii*). Therefore, this gene may have been involved in the cognitive evolution of multiple *Hominidae* lineages, including our own.

Several additional prominent candidates of positive selection are involved in cognitive and neurodevelopmental phenotypes. This includes *AUTS2*, identified by ELS in *G. g. gorilla* (second highest rank) and *Pan troglodytes troglodytes* (fourth highest rank), and implicated in neuronal development and autism in humans (Oksenberg and Ahituv 2013). In addition, *CSMD1*, the gene with FWH signatures in the most lineages (*Homo sapiens*, *Pan paniscus*, *Pan troglodytes ellioti*, *Pan troglodytes schweinfurthii*, *Gorilla gorilla gorilla* and *Pongo pygmaeus*) (supplementary table S73, Supplementary Material online), whose function is unknown but that is highly expressed in the central nervous system (Kraus et al. 2006) and harbors variants associated with schizophrenia (Håvik et al. 2011). Further, of the four genes with FWH signatures in five lineages, two are associated with neuronal phenotypes: *KCNIP4*, which encodes an A-type potassium channel modulatory protein, and harbors variants associated with attention deficit hyperactive disorder (Weißflog et al. 2013) and

NRG3 (Neuregulin 3), which is crucial in the development of the nervous system and whose variants are associated to schizophrenia (Chen et al. 2009).

In addition, 12 genes detected as positively selected by MK are related to neurodevelopmental disorders in humans (supplementary table S98, Supplementary Material online). Five (*MCPH1*, *CASC5*, *PHGDH*, *FTO* and *NBN*) can display a phenotype of microcephaly when mutated (Faheem et al. 2015), with mutations in *MCPH1* and *CASC5* being responsible for autosomal recessive primary microcephaly (MCPH) (Woods et al. 2005; Genin et al. 2012). *MCPH1* (identified here in *H. sapiens*) has been described as a target of positive selection in primate evolution (Wang and Su 2004; Shi et al. 2013); *CASC5* (identified here in *Pan troglodytes ellioti* and *Pan paniscus*) contains, in *Homo sapiens*, a nonsynonymous mutation that reached fixation since the split with Neandertals (Prüfer et al. 2014), suggesting recent positive selection also in our lineage. *MCPH1*. *CENPJ*, another MCPH gene, shows marginally nonsignificant evidence of positive selection (P value = 0.055 in *P.t. verus*). Together these results show putative adaptive evolution in genes that may have contributed to changes in brain size and function during primate evolution.

Conclusion

We present a comparative population genomic analysis that investigates the influence of natural selection across the *Hominidae*. This information sheds light on the past adaptations of each of these populations. As expected, immune function was a strong selective force in all species. Given the close evolutionary relationship, similar physiology and shared pathogens of humans with the other *Hominidae* lineages, further functional study of these immunity-related genes may be of medical relevance. In addition, the evidence of positive selection in neuronal pathways of several lineages suggests differential adaptations in phenotypes that distinguish the *Hominidae* species from one another. For example, genes that show strong signals of positive selection solely on the human lineage constitute the best candidates to explain human-specific neurological phenotypes. Similarly, genes with evidence of positive selection in species that differ from one another in phenotypes including size, locomotion, morphology or diet help us to understand the genetic basis of these adaptations.

The fact that even the modest differences in long-term N_e between *Hominidae* lineages has had discernible impacts on the efficacy of natural selection, both to remove deleterious alleles and to favor adaptive ones, has additional implications. The different great ape species, all of which (except for humans) are currently endangered, may thus differ significantly in their ability to adapt to environmental change. This may affect their ability to adapt not only to constantly changing pathogens, but also to the often human-induced changes to their habitats.

Methods

Dataset

The dataset we analyzed consists of whole-genome autosomal sequences from 83 individuals across all the major

lineages of the *Hominidae* (with the exception of *Gorilla beringei beringei*) (fig. 1 and supplementary table S1, Supplementary Material online). The dataset was originally presented in Prado-Martinez et al. (2013; SOM), where the SNP calling pipeline and filtering criteria are described in detail. All reads are mapped to the human reference genome (hg18). This approach has three main advantages. First, we take advantage of the extensive data-quality exploration and filtering performed in the original publication. Second, mapping to the human genome ensures that all species are mapped to a high-quality genome, avoiding the (hard to account for) biases that would result from mapping to genomes of low and varying qualities. Third, the human genome has the most comprehensive annotation of gene coding regions, which is very important in this study.

To avoid errors introduced by miss-mapping due to paralogous variants, we also restricted all analyses to a set of sites with a unique mapping to the human genome. To address the possible influence of unknown copy number variants (which would result in collapsing several genomic regions during mapping and produce false SNP calls), we took several steps (supplementary fig. S1, Supplementary Material online). Using UCSC tracks we excluded from analysis all repetitive regions (~248 Mb), segmental duplications (~154 Mb), genomic gaps (~226 Mb) and tandem repeats (~38 Mb). We also excluded structural variants detected in any of the great ape lineages (~334 Mb) based on the most comprehensive catalogue available which was itself generated using this dataset and read-depth methods (Sudmant et al. 2013). Furthermore, sites with depth of coverage (DP) < (mean read depth/8.0) and DP > (mean read depth × 3), were also removed. To maximize the number of sites to be analyzed, we excluded multiple individuals with low coverage (supplementary tables S1 and S2, Supplementary Material online). Additionally, we also required positions to have at least 5× coverage in all individuals per species. Only the resulting set of sites, which we termed “callable sites”, were used in further analyses; this minimizes, as much as possible, the effects of filtering in all enrichment analyses. This resulted in a mean of 2,099 Mb of analyzable genome sequence per species (supplementary fig. S1, Supplementary Material online). We caution that despite our many efforts, which include multiple stringent filtering steps and the manual curation of targets presented in the main text, we cannot discard the presence of some artifact in our data (e.g., undetected structural variants in the candidate targets of balancing selection) although we expect that them to have a weak influence in our overall results.

Tests

Hudson–Kreitman–Aguadé Test (HKA)

To detect long-term balancing selection and positive selection that could have occurred at a deep evolutionary time-scale, we used a statistic based on the HKA test (Hudson et al. 1987). Here, the HKA statistic is simply the ratio of polymorphic (SNPs) to divergent (substitutions) sites in a window. We consider as a polymorphism a genomic position that was identified as a single nucleotide variant (SNV) in Prado-Martinez et al. (2013). We consider a substitution (a divergent

site) a genomic position that is identified as a fixed difference between the tested and the outgroup lineage. For consistency, *Homo sapiens* was used as an outgroup for all lineages. When performing the test for *Homo sapiens*, we used the combined *Pan troglodytes* lineages as the outgroup.

For each lineage, the genome was divided into 30-kb genomic windows with 15-kb overlap and the HKA statistic was calculated. We consider only windows that contain at least 300 callable and 6 informative sites, where an informative site is a SNV or substitution (see supplementary materials HKA, Supplementary Material online). Each window in the genome was ranked according to its HKA score, and the rank was considered the window’s empirical *P* value (see supplementary fig. S4, Supplementary Material online, for an example of the distribution of polymorphic sites and substitutions across the HKA empirical distribution). To ensure that our results were not influenced by variation in data quality across the genome, we tested whether extreme HKA scores are biased in terms of coverage or mapping quality. We find no evidence for such artifacts influencing our results (see supplementary materials HKA and supplementary figs. S2 and S3, Supplementary Material online).

Fay and Wu *H* Test (FWH)

To detect complete or nearly complete positive selective sweeps caused by recent or ongoing positive selection, which result in an excess of high-frequency derived alleles, we used the FWH statistic (Fay and Wu 2000). We confirmed our implementation of the FWH statistic was capable of detecting recent selective sweeps using simulations (see supplementary materials FWH 1 and supplementary fig. S19, Supplementary Material online). For each lineage, the genome was divided into 30-kb windows with 15-kb overlap using the same strategy as the HKA test (see above and supplementary materials FWH 1, Supplementary Material online). Windows with less than 300 callable sites were removed. Each window in the genome was ranked according to its FWH score, and the rank was considered the window’s empirical *P* value.

McDonald–Kreitman Test (MK)

To detect positive and purifying selection on protein coding genes, we used the McDonald–Kreitman test (McDonald and Kreitman 1991). The MK test was calculated for all lineages with at least five individuals, as this was considered the minimum sample size for sufficient polymorphism data (supplementary table S93, Supplementary Material online). Only *Pan troglodytes troglodytes* did not meet these criteria. *Pan t. verus* met the criterion only by including Donald, an individual excluded in all other analyses because of evidence of admixture between *P.t. verus* and *Pan troglodytes troglodytes* (Prado-Martinez et al. 2013). Coordinates for coding regions of all autosomal transcript unique identifiers were taken from RefSeq hg18 and intersected with the callable sites in our data (Pruitt et al. 2012). This resulted in ~15.1 Mb of coding sequence available for analysis. For each lineage, we count all polymorphisms and substitutions that are predicted to have appeared after the most recent common ancestor with an

outgroup (*Homo sapiens* was used for all lineages, except when performing the test for *Homo sapiens*, in which case *Pan troglodytes* was used) (supplementary table S94, Supplementary Material online). The total number of transcripts tested for each species can be seen in supplementary table S6C, Supplementary Material online, and the significant transcripts for either positive or purifying selection in supplementary tables S96 and S97, Supplementary Material online. Variants were annotated as either synonymous or nonsynonymous using ANNOVAR (Wang et al. 2010). Multiallelic sites were excluded (see supplementary materials MK 1.4, Supplementary Material online).

Extended Lineage Sorting Test (ELS)

To detect lineage-specific positive selection that occurred after the divergence of two closely related lineages, we scan the genome for a signal of extended lineage sorting (see SOM 13 in Green et al. 2010; see Supplementary Information 7 in Prüfer et al. 2012; see Supplementary Information 19a in Prüfer et al. 2014), i.e., genomic regions where the lineage of a closely related outgroup falls basal to the lineages of a test-group of individuals. The test requires a particular relationship between the test-group and the closely related outgroup individual where the outgroup individual is sufficiently close and the test-group is sufficiently diverse so that the outgroup often falls within the diversity of the test-group. To determine which population pairs are suitable for ELS, we performed neutral coalescent simulations with ms (Hudson 2002) (supplementary table S66, Supplementary Material online). The fraction of derived sites in the simulations was compared with the fraction in the data, which closely matched the simulations in most cases (supplementary fig. S20 and supplementary tables S66 and S67, Supplementary Material online). Three lineage pairs showed a sufficiently close relationship and were used for the ELS test: *Pan troglodytes*—*Pan paniscus*, *Gorilla g. gorilla*—*Gorilla b. graueri*, *Pongo abelii*—*Pongo pygmaeus*.

We use an implementation of the ELS test that is based on a hidden Markov model (HMM) that analyses SNPs in individuals from one population and the genotype from a single individual from the outgroup population (Prüfer et al. 2014, SOM). The HMM then infers the posterior probability for the hidden states *internal* (the outgroup falls within the diversity of the test group) and *external* (the outgroup falls basal to the lineages of the test group) at all SNP positions.

Following Prüfer et al. (2012, Supplementary Information 7), external regions were defined as a run of SNPs with a probability of >0.8 for being external that is not interrupted by SNPs with a probability of >0.8 for being internal, and scored by their genetic length using the 1-Mb average human recombination rate from Kong et al. (2002).

For each population, the HMM was run repeatedly with each “outgroup” individual. To combine these multiple outputs, we disregarded any external region that was not in the top 5% of the empirical distribution in all runs (as truly external regions are shared among all outgroup individuals) and the remaining external regions were then assigned a final rank

based on their cumulative rank score from the multiple runs (supplementary materials ELS 1.3 and supplementary tables S68–70, Supplementary Material online).

Region Annotation and GO Category Enrichment

Regions were annotated as genic (protein-coding and nonprotein coding) if at least 1 bp of the region overlapped with a gene using GENCODE hg18 gene coordinates (Harrow et al. 2012).

To test for evidence of functional enrichment among the genes that we detect as putative targets of natural selection, we performed biological category enrichment analysis using the software WebGestalt (Zhang et al. 2005). For the HKA and FWH tests, we selected the 200 genes with the strongest signatures of selection as our test set of candidate genes. For the ELS test, we considered all genes in the 5% longest external regions. For the MK test, we selected by species all genes with at least one transcript presenting a nominal P value of ≤ 0.05 .

For each test and lineage, we tested for functional enrichment using several databases of biological pathway and functional information: GO categories (Harris et al. 2004) “biological processes”, “molecular functions” and “cellular components”; the Kyoto encyclopedia of genes and genomes (KEGG) pathway database (Kanehisa et al. 2004) based on mammalian and human phenotype ontology; and the PheWas database, which is based on the human PheWas ontology (Denny et al. 2010). We set a significance threshold of 0.05 and used the Bonferroni correction for multiple hypothesis testing. Significant categories driven by only one gene were discarded due to the high potential for spurious signals in such cases. For HKA results, see supplementary tables S3N–S3AAA, Supplementary Material online, for ELS see supplementary tables S71 and S72, Supplementary Material online, for FWH see supplementary tables S5C–S5N, Supplementary Material online, for MK test see supplementary table S101, Supplementary Material online.

We note that all gene pathways used were annotated for humans. While this is not ideal for pathway enrichment analyses of nonhuman species, the putative biases should be minor. This is because these functional elements are evolutionarily constrained and these species are extremely closely related (all within only 12 My). For example, there have only been 96 gene-deletion events in the great apes (Prado-Martinez et al. 2013), which should have a minimal impact on an enrichment analyses that uses thousands of genes. Furthermore, any putative annotation errors between species should be random with respect to biological pathways and not systematically biasing gene enrichment results.

We tested the potential effect of gene length bias on the results by repeating the enrichment analyses after randomly selecting equal numbers of windows and exploring the overlap of these (random) categories with our results (see supplementary materials HKA 7, Supplementary Material online).

Data Access

An interactive browser with the signatures of natural selection for each species is available at <http://tinyurl.com/nf8qmzh> (last accessed October 10, 2016).

Supplementary Material

Supplementary figures S1–S22, tables S1–S107 and supplementary materials are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

We thank Matthias Ongyerth for assistance with data preparation, Michael Lachmann and Mark Stoneking for discussions and valuable comments, and Michael Lachmann for help with the ELS test. We thank the members of the Great Ape Genome Diversity Consortium for support throughout this work. This work was supported by funding from the Max Planck Society to K.P. and A.M.A.; by grants from the Ministerio de Economía y Competitividad in Spain (grant BFU2013-43726-P) and the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya (grant GRC 2014 SGR 866) to J.B.; by an European Research Council Advanced Grant (233297) to S.Pääbo and European Research Council Starting Grant (260372) to T.M.B.; and by European Molecular Biology Organization Young Investigator Award and Ministerio de Ciencia e Innovación in Spain (BFU2014-55090-P) to T.M.B.

References

- Anderson RM, May RM. 1982. Coevolution of hosts and parasites. *Parasitology* 85(02):411–426.
- Andrés AM, Dennis MY, Kretzschmar WW, Cannons JL, Lee-Lin S-Q, Hurle B, Schwartzberg PL, Williamson SH, Bustamante CD, Nielsen R, et al. 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet.* 6:e1001157.
- Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, Clark AG. 2009. Targets of balancing selection in the human genome. *Mol Biol Evol.* 26(12):2755–2764.
- Bataillon T, Duan J, Hvilsom C, Jin X, Li Y, Skov L, Glemin S, Munch K, Jiang T, Qian Y, Hobolth A. 2015. Inference of purifying and positive selection in three subspecies of chimpanzees (*Pan troglodytes*) from exome sequencing. *Genome Biol Evol.* 7(4):1122–1132.
- Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441(7089):87–90.
- Boasso A, Shearer GM. 2008. Chronic innate immune activation as a cause of HIV-1 immunopathogenesis. *Clin Immunol.* 126:235–242.
- Boasso A, Vaccari M, Fuchs D, Hardy AW, Tsai W-P, Trynieszewska E, Shearer GM, Franchini G. 2009. Combined effect of antiretroviral therapy and blockade of IDO in SIV-infected rhesus macaques. *J Immunol.* 182:4313–4320.
- Bustamante CD, Fedel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sniinsky JJ, Hernandez RD, Civallo D. 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153–1157. 20
- Cagliani R, Riva S, Biasin M, Fumagalli M, Pozzoli U, Lo Caputo S, Mazzotta F, Piacentini L, Bresolin N, Clerici M, et al. 2010. Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection. *Hum Mol Genet.* 19:4705–4714.
- Cagliani R, Riva S, Pozzoli U, Fumagalli M, Comi GP, Bresolin N, Clerici M, Sironi M. 2011. Balancing selection is common in the extended MHC region but most alleles with opposite risk profile for autoimmune diseases are neutrally evolving. *BMC Evol Biol.* 11(1):171.
- Casals F, Sikora M, Laayouni H, Montanucci L, Muntasell A, Lazarus R, Calafell F, Awadalla P, Netea MG, Bertranpetit J. 2011. Genetic adaptation of the antibacterial human innate immunity network. *BMC Evol Biol.* 11:202.
- Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet.* 10:195–205.
- Chen PL, Avramopoulos D, Lasseter VK, McGrath JA, Fallin MD, Liang KY, Nestadt G, Feng N, Steel G, Cutting AS, Wolyniec P. 2009. Fine mapping on chromosome 10q22-q23 implicates Neuregulin 3 in schizophrenia. *Am J Hum Genet.* 84:21–34.
- Coop G. 2016. Does linked selection explain the narrow range of genetic diversity across species? *bioRxiv* 1:042598.
- Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. “The role of geography in human adaptation.”. *PLoS Genet.* 5:e1000500.
- Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol.* 13:e1002112.
- Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 26:1205–1210.
- Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci.* 110(14):5294–5300.
- Ellegren H, Galtier G. 2016. Determinants of genetic diversity. *Nat Rev Genet* 17:422–433.
- Elyashiv E, Bullaughey K, Sattath S, Rinott Y, Przeworski M, Sella G. 2010. Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. *Genome Res.* 20:1558–1573.
- Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. *Genome Res.* 24:885–895.
- ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9(4):e1001046.
- Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol.* 21:569–575.
- Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol.* 26:2097–2108.
- Faheem M, Naseer MI, Rasool M, Chaudhary AG, Kumosani TA, Ilyas AM, Pushparaj P, Ahmed F, Algahtani HA, Al-Qahtani MH, et al. 2015. Molecular genetics of human primary microcephaly: an overview. *BMC Med Genomics* 8(Suppl 1):S4.
- Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.
- Ferrer-Admetlla A, Bosch E, Sikora M, Marqués-Bonet T, Ramírez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, et al. 2008. Balancing selection is the main force shaping the evolution of innate immunity genes. *J Immunol.* 181:1315–1322.
- Genin A, Desir J, Lambert N, Biervliet M, Van der Aa N, Pierquin G, Killian A, Tosi M, Urbina M, Lefort A, et al. 2012. Kinetochore KMN network gene CASCS mutated in primary microcephaly. *Hum Mol Genet.* 21:5306–5317.
- Gokcumen O, Tischler V, Tica J, Zhu Q, Iskow RC, Lee E, Fritz MH, Langdon A, Stütz AM, Pavlidis P, Benes V. 2013. Primate genome architecture influences structural variation mechanisms and functional consequences. *Proc Natl Acad Sci.* 110(39):15764–15769.
- Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328:710–722.
- Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al. 2013. Identifying recent adaptations in large-scale genomic data. *Cell* 152:703–713.
- Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, Adams DJ, Keightley PD. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet.* 5:e1003995.
- Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32:D258–D261.

- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* 22:1760–1774.
- Håvik B, Le Hellard S, Rietschel M, Lybæk H, Djurovic S, Mattheisen M, Mhleisen TW, Degenhardt F, Priebe L, Maier W, et al. 2011. The complement control-related genes CSMD1 and CSMD2 associate to schizophrenia. *Biol Psychiatry.* 70:35–42.
- Hedrick PW. 1999. Balancing selection and MHC. *Genetica* 104:207–214.
- Heeney J, Jonker R, Koomstra W, Dubbes R, Niphuis H, Di Rienzo AM, Gougeon ML, Montagnier L. 1993. The resistance of HIV-infected chimpanzees to progression to AIDS correlates with absence of HIV-related T-cell dysfunction. *J Med Primatol.* 22:194–200.
- Hubisz MJ, Pollard KS. 2014. Exploring the genesis and functions of human accelerated regions sheds light on their role in human evolution. *Curr Opin Genet Dev.* 29:15–21.
- Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
- Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.
- Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet.* 32:415–435.
- Irvine AD, McLean WHI. 2006. Breaking the (un)sound barrier: filaggrin is a major gene for atopic dermatitis. *J Invest Dermatol.* 126:1200–1202.
- Jensen JD, Bachtrog D. 2011. Characterizing the influence of effective population size on the rate of adaptation: Gillespie's Darwin domain. *Genome Biol Evol.* 3:687–701.
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32:D277–D280.
- Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* 16:980–989.
- Key FM, Fu Q, Romagné F, Lachmann M, Andrés AM. 2016. Human adaptation and population differentiation in the light of ancient genomes. *Nat Commun.* 18:7.
- Key FM, Teixeira JC, Filippo C, de Andrés AM. 2014. Advantageous diversity maintained by balancing selection in humans. *Curr Opin Genet Dev.* 29:45–51.
- Kimura M. 1979. The neutral theory of molecular evolution. *Sci Am.* 241:98–126.
- King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–116.
- Kong A, Gudbjartsson DF, Sainz J, Jonsson GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet.* 31:241–247.
- Kraus DM, Elliott GS, Chute H, Horan T, Pfenninger KH, Sanford SD, Foster S, Scully S, Welcher AA, Hokers VM. 2006. CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. *J Immunol.* 176:4419–4430.
- Lee Y, Langley C, Begun D. 2014. Differential strengths of positive selection revealed by hitchhiking effects at small physical scales in *Drosophila melanogaster*. *Mol Biol Evol* 31(4):804–816.
- Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species?. *PLoS Biol.* 10(9):e1001388.
- Lewontin RC. 1974. The genetic basis of evolutionary change. New York: Columbia University Press.
- Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, De Vos M, Dixon A, Demarche B. 2007. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* 3(4):e58.
- Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang SP, Wang Z, Chinwalla AT, Minx P, Mitreva M. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* 469(7331):529–533.
- McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654.
- McGrew WC. 2004. The cultured chimpanzee. Reflections on cultural primatology. Cambridge University Press.
- McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, Ryder OA, Ape Genome Project G, Kidd JM, Wall JD, et al. 2015. Inference of gorilla demographic and selective history from whole-genome sequence data. *Mol Biol Evol.* 32:600–612.
- McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, Boerwinkle E. 2007. A common allele on chromosome 9 associated with coronary heart disease. *Science* 316(5830):1488–1491.
- McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* 5(5):e1000471.
- Metz R, Smith C, DuHadaway JB, Chandler P, Baban B, Merlo LMF, Pigott E, Keough MP, Rust S, Mellor AL, et al. 2014. *IDO2* is critical for *IDO1*-mediated T-cell regulation and exerts a non-redundant function in inflammation. *Int Immunol.* 26:357–367.
- Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang S-P, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK, et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.
- Murray MF. 2010. Insights into therapy: tryptophan oxidation and HIV infection. *Sci Transl Med.* 2:32ps23.
- Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, Indap A. 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* 19(5):838–849.
- Oji V, Eckl KM, Aufenvenne K, Natebus M, Tarinski T, Ackermann K, Seller N, Metzke D, Nurnberg G, Folster-Holst R, et al. 2010. Loss of corneodesmosin leads to severe skin barrier defect, pruritus, and atopy: unraveling the peeling skin disease. *Am J Hum Genet.* 87(2):274–281.
- Oksenberg N, Ahituv N. 2013. The role of *AUTS2* in neurodevelopment and human evolution. *Trends Genet.* 29:600–608.
- Pagel M, Meade A. (2013). BayesTraits V2. Software and manual. Reading: University of Reading. <http://www.evolution.rdg.ac.uk/BayesTraitsV2Beta.html>.
- Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Res.* 21(11):1769–1776.
- Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* 499:471–475.
- Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20(4):R208–R215.
- Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486:527–531.
- Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43–49.
- Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 2(40):D130–D135.
- Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreño-Torres A, Pavlidis P, Laayouni H, Bertranpetit J, Engelken J. 2014. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.* 42 (Database issue):D903–D909.
- Pybus M, Luisi P, Dall'Olio G, Uzkudun M, Laayouni H, Bertranpetit J, Engelken J. 2015. Hierarchical boosting: a machine-learning

- framework to detect and classify hard selective sweeps in human populations. *Bioinformatics* 31(24):3946–3952.
- Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* 312 (5780):1614–1620.
- Scally A, Dutheil J, Hillier L. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483:169–175.
- Semendeferi K, Lu A, Schenker N, Damasio H. 2002. Humans and great apes share a large frontal cortex. *Nat Neurosci.* 5:272–276.
- Shi L, Li M, Lin Q, Qi X, Su B. 2013. Functional divergence of the brain-size regulating gene MCPH1 during primate evolution and the origin of humans. *BMC Biol.* 11:62.
- Smith FJD, Irvine AD, Terron-Kwiatkowski A, Sandilands A, Campbell LE, Zhao Y, et al. 2006. Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. *Nat Genet.* 38:337–342.
- Südhof TC. 2008. Neuroligins and neuroligins link synaptic function to cognitive disease. *Nature* 455:903–911.
- Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* 23(9):1373–1382.
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, Konkel MK. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526(7571):75–81.
- Tomasello M, Call J. 1997. Primate cognition. USA: Oxford University Press
- Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164.
- Wang YQ, Su B. 2004. Molecular evolution of microcephalin, a gene determining human brain size. *Hum Mol Genet.* 13:1131–1137.
- Weißflog L, Scholz CJ, Jacob CP, Nguyen TT, Zamzow K, Groß-Lesch S, Renner TJ, Romanos M, Rujescu D, Walitza S, et al. 2013. *KCNIP4* as a candidate gene for personality disorders and adult ADHD. *Eur Neuropsychopharmacol.* 23:436–447.
- Woods CG, Bond J, Enard W. 2005. Autosomal recessive primary microcephaly (MCPH): a review of clinical, molecular, and evolutionary findings. *Am J Hum Genet.* 76:717–728.
- Zhai W, Nielsen R, Slatkin M. 2009. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Mol Biol Evol.* 26(2):273–283.
- Zhang B, Kirov S, Snoddy J. 2005. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* 1(33):W741–W748.

Chapter 3

Additional Supplementary figures S7-S15 and Supplementary tables S1-107 are found in USB stick accompanying this thesis.

Supplementary Material

Natural Selection in the Great Apes

Alexander Cagan, Christoph Theunert, Hafid Laayouni, Gabriel Santpere, Marc Pybus, Ferran Casals, Kay Prüfer, Arcadi Navarro, Tomas Marques-Bonet, Jaume Bertranpetit, Aida M. Andrés

Section 1: Sample processing	2
Section 2: HKA test	4
Section 3: FWH test	19
Section 4: MK test.....	22
Section 5 : ELS test.....	32
Section 6 : Subsampling analysis.....	37
Section 7: Targets of selection.....	40

Section 1: Sample processing

Gabriel Santpere, Alexander Cagan

1. Samples

1. Samples

The dataset we analyzed consists of whole-genome autosomal sequences from 83 individuals across all the major lineages of the *Hominidae* (with the exception of *G.b. beringei*) (Figure 1 & supplementary table S1). The dataset was originally presented in Prado-Martinez et al. (Prado-Martinez et al. 2013; SOM), where the SNP calling pipeline and filtering criteria are described in detail. All reads are mapped to the human reference genome (hg18). This approach has three main advantages. First, we take advantage of the extensive data-quality exploration and filtering performed in the original publication. Second, mapping to the human genome ensures that all species are mapped to a high-quality genome, avoiding the (hard to account for) biases that would result from mapping to genomes of low and different qualities. Third, the human genome has the most comprehensive annotation of gene coding regions, which is a central object of this study.

1.1. Filtering strategy

Genomic data is prone to many artifacts that can bias downstream analyses. To account for this we applied a comprehensive series of filters which are summarized in supplementary fig S1. To avoid errors introduced by miss-mapping caused by multi-copy sequences or structural variants among species, we restricted all analyses to genomic intervals with a unique mapping to the human genome. Using UCSC tracks we excluded from analysis all repetitive regions identified by RepeatMasker (~248 Mb), if repeat divergence were lower than 10%, and from Tandem Repeat Finder (~38 Mb). In both cases we only masked repeats longer than 80 bp, since repeats shorter than reads can be better mapped by their flanking non-repetitive sequence. We also masked segmental duplications (~154 Mb) and genomic gaps (~226 Mb). Finally, we also excluded structural variants detected in any of the great ape lineages (~334 Mb). These regions were detected in a previous study of structural variation in great apes which analysed this dataset (Sudmant et al., 2013).

Exclusion of regions that may contain unannotated structural variants removes a potential source of false positives in our analyses. We expect filtering out structural variants to have a minor effect on our subsequent analyses. As our approach is empirical (except for the MK, where we compare synonymous and non-synonymous mutations and thus removing CNVs would have minimal effect) we do not attempt to estimate the proportion of the genome under natural selection. We expect the tails of the empirical

distributions of our selection tests to be enriched in true candidates of natural selection regardless of the removal of CNVs.

Furthermore, sites with depth of coverage (DP) < (mean read depth/8.0) and DP > (mean read depth * 3), were also removed. To maximize the number of sites to be analyzed we excluded multiple individuals with low coverage (supplementary tables S1 and S2). Some variable sites presented missing data introduced by an Allele Balance Filter used in Prado-Martinez et al. to account for the putative contribution of traces of genetic contamination. These positions were generally excluded particularly in analysis dependent on allele frequencies. Additionally, we also required positions to have at least 5x coverage in all individuals per species. To increase the number of callable sites we excluded low coverage individuals. Only the resulting set of sites, which we termed 'callable sites', were used in further analysis. This resulted in a mean of 2,099 Mb of analyzable genome sequence per species (supplementary fig S1).

References

- Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* **499**: 471–5.
- Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome research* Sep 1;23(9):1373-82.

Section 2: HKA test

Alexander Cagan, Christoph Theunert, Aida M. Andrés

1. The HKA test

We aim to identify candidate targets of natural selection by analyzing the patterns of diversity and divergence in the genomes of each lineage. The neutral model predicts that the ratio of polymorphism to divergence will be approximately constant across the genome, and regions where this ratio is highly skewed in either direction represent potential deviation from neutrality due to selection. The Hudson-Kreitman-Aguade (HKA) test is a neutrality test that contrasts intra-population diversity to inter-populations divergence in a test locus, and compares it with putatively neutral loci (Hudson et al. 1987). By contrasting patterns of diversity and divergence in genomic windows without any differentiation of loci by their function (e.g. coding and non-coding regions), the HKA test is able to detect signatures of selection without an *a priori* assumption that selection will be in a particular type of loci (e.g. coding regions). The ratio of diversity to divergence also allows the HKA test to detect selection at a greater time scale than many SFS-based tests, which lose power once the selected locus has recovered to equilibrium levels of diversity. A previous study comparing the statistical power of selection tests found the HKA test to have the most power to detect positive selection (Zhai et al. 2009). At the genome scale we apply a test based on the HKA by analyzing windows across the genome to identify loci that are outliers of the empirical distribution, either because their ratio of polymorphisms to substitutions is too high or because it is too low when compared with the rest of the genome; these regions represent candidate targets of natural selection in the great apes. Because different types of selection affect differently the ratio of polymorphism to divergence, this test allows us to distinguish among different types of selection. Regions with the lowest HKA scores, where diversity levels are lowest relative to divergence, are candidates for being under positive or negative selection (diversity-reducing types of selection). Regions of the genome where the HKA score is highest, due to high levels of diversity relative to divergence, are suggestive of long-standing balancing selection. Regions of the genome in the middle part

of the empirical distribution are expected to be either neutral or subject to varying levels of purifying selection.

When whole-genome data is unavailable, significance of the HKA test has been estimated based on neutral coalescent simulations, which allow us to predict the range of HKA scores expected under neutrality. Here, with availability of full genome data, and under the reasonable assumptions that the majority of the genome is evolving neutrally or under purifying selection, and that a small portion of the genome is evolving under positive and balancing selection, candidate targets of natural selection can be obtained from the tails of the empirical distribution. Regions of the genome with scores at either extremes of the HKA score empirical distribution have the most unusual ratios of polymorphism to divergence, and are thus potential targets of natural selection.

1. Overview of steps taken to avoid and account for possible artifacts

An unusual level of polymorphism may result from technical artifacts and thus not represent true signatures of natural selection. We address this problem by taking several steps to ensure signals are not the result of technical artifacts. Particular attention was taken to artifacts that may result in increased polymorphism, since the HKA is the only test in this study that aims to identify the signatures of balancing selection, but our filtering strategy should be effective in removing artifacts that may result both in high and in low levels of diversity. We present here a summary of the steps we followed to ensure that we detect true biological signals rather than technical artifacts; details can be found in the sections below.

1.1. Pre-test Filtering

The fact that not all populations are analyzed for exactly the same sites can also lead to unequal power for the HKA test. To reduce the effect of these possible biases, data was filtered using the following criteria: for each pair-wise HKA test sites were excluded if they had a read depth of $< 5x$ in any individual in either of the two populations analysed.

1.2. Performing the HKA test

The test, based on the HKA, was calculated in 30kb windows with a 15kb overlap between windows across the genome. For each window and population, we computed the number of polymorphic sites and the number of fixed substitutions to the out-group. Fixed substitutions are positions where both populations are fixed homozygous for alternative alleles. Here, our HKA score for each window was calculated by dividing the number of polymorphic sites (diversity) by the number of fixed substitutions (divergence). Windows with less than 300 callable sites or with less than 6 informative sites (substitutions between populations or within-populations polymorphisms) were removed from the analysis as the scarcity of sites may introduce too much noise. *H. sapiens* was used as an out-group for all tests for consistency, and *P. troglodytes* was used as an out-group for analyzing *H. sapiens*. In *P. troglodytes*, the test was carried out at the sub-populations level, for each of the four *P. troglodytes* sub-populations separately and with *H. sapiens* as out-group. After performing the HKA test we confirm that the regions in the extreme tails of the HKA score distribution are not unusual in terms of their coverage, mapping quality, or the number of duplications they contain (which addresses possible mapping problems as well as gene conversion). We find no systematic bias in any of these categories (see Supplementary Materials HKA 1.3 and 1.4), showing that technical artifacts do not substantially contribute to the regions we detect as candidates of natural selection.

1.3. Coverage Bias

One possible source of unusual diversity levels is extremely high or low coverage in the analyzed window. To ensure that this technical aspect of the data is not substantially contributing to the extremes of the HKA distribution (so we detect true biological signals rather than technical artifacts), we tested whether extreme HKA scores are biased towards lower or higher average coverage than other regions of the genome. To do this, we divided the 30kb windows which are the output of the HKA test into three categories based on their HKA score: regions with HKA scores in the top 1%, bottom 1% and middle 98% of the empirical distribution.

We also ran permutations from each of the three categories to ensure that there is no bias that is being obscured because the middle 98% of the distribution has a much larger number of data points than the top and bottom 1% distributions. For each permutation, 50 regions were randomly selected from each of the distributions and their average coverage was calculated. This was repeated 1000 times with replacement. These results show no evidence of differences in coverage among the groups of loci with average and with extreme HKA scores (supplementary fig S2 for an example with *G.g. gorilla*).

1.4. Mapping Quality Bias

To ensure that mapping problems did not affect our analysis, we calculated the average Mapping Quality score across all callable sites for the 30kb windows in the bottom 5%, top 5%, and middle 90% of the genomic distribution of HKA scores (see supplementary fig S3 for an example with *G.g. gorilla*). All regions have an average Mapping Quality Score within the range of 56-59, well above the mapping quality threshold of 25 used for the dataset, and with no skew towards lower Mapping Quality in the top tail of the distribution, suggesting no systematic bias towards low mapping scores in the extreme tails of the HKA distribution.

2. Percentage of windows containing exons analysis

To explore the results of the HKA test and to investigate the power of the test to detect regions under natural selection we tested if there was an enrichment of windows containing functional elements in the tails of our HKA score distribution. As most of the genome is believed to be non-functional but natural selection targets functional sites, we expect an enrichment of windows containing functional elements in the tails of the HKA empirical distribution.

2.1. Methods

To test this we first created two different lists of functional sites based on the GENCODE annotation of the *H. sapiens* genome. We made one list with the start and end coordinates of all exons classified as being

from protein-coding genes. We made another list with the start and end coordinates of all exons classified as being from either RNA (mtRNA , miRNA, Mt_rRNA, misc_RNA, rRNA, snRNA, snoRNA) or non-protein-coding genes. This allowed us to additionally test whether there was a difference in the pattern of HKA scores between protein-coding compared to non-protein-coding regions.

For each list separately, and for each lineage, we annotated the 30kb HKA windows as overlapping (or not overlapping) with an exon by at least 1bp. We then defined different thresholds of the tail of the empirical distribution of the HKA score, and calculated what percentage of windows within each percentile overlap with exons.

Results

2.2. Protein-coding exons

The results for protein-coding exons show that there is a trend for the number of windows containing protein-coding exons to increase when moving from windows with high HKA scores to windows with low HKA scores, as long as the windows are not in the far tails of the empirical distribution (e.g. from the 99-5% bins, Figure 2 and supplementary table S3). As the lower HKA scores tend to be driven by reduction in diversity rather than an excess of divergence (supplementary fig S4) this suggests that this trend of exonic enrichment towards lower HKA scores is most likely driven by the effect of purifying selection, which are on average stronger in exonic regions of the genome (McVicker et al. 2009).

It is interesting that at the very bottom tail of the HKA empirical distribution all lineages show a slight difference in trends, with the percentage of windows containing protein-coding exons not always increasing further as we move further into the tail (depending on the populations at different points between the bottom 5% and the end of the bottom tail). There could be several reasons for this change, which is observed in all lineages. One explanation could be noise, as the number of regions decreases when moving further in the tail. An alternative explanation is that the most strongly selected targets of selection are non-coding variants in regulatory regions that are not close to genes. A third explanation is that some selective sweeps targeting exonic regions extend far beyond the exons themselves, resulting in an excess of windows

in the extreme bottom tail of the HKA empirical distribution that do not contain exons but have low diversity because they are part of a selective sweep on an exonic variant. This is not unexpected as fast selective sweeps can extend very long genomic regions because the short-term effects of recombination do not break the association across variants (Maynard Smith & Haigh, 1974). In fact, we observe an unusually high clustering of windows among those present in the bottom tail of the HKA empirical distribution (see Supplementary Materials HKA 4)

It is also interesting that in the extreme top tail of the HKA empirical distribution (candidate regions under balancing selection) we observe a change with respect to the overall trend described above. Specifically, we observe a drastic increase in the percentage of windows overlapping exons for the windows more strongly enriched in SNPs, in all the populations. This is unexpected under neutral evolution and under purifying selection, and is likely to be the result of balancing selection acting on or near these protein-coding exons. Also, the signature of long-term balancing selection is narrow due to the long-term effects of recombination (e.g. Kaplan et al. 1988; Charlesworth et al. 1997) and we do not expect several windows to show the signatures of one single event. This conclusion is further supported by the observation that in all lineages many of these genes are from the MHC region, a well-known target of balancing selection across vertebrates (Hughes et al. 1998). Therefore this analysis provides evidence to support the idea that long-term balancing selection is not prevalent across the genome (although we note that we may have limited power to detect its signatures in 30kb windows). Despite the small number of regions, many of them are shared across populations, emphasizing the important and conserved role that balancing selection plays in maintaining adaptive diversity, particularly with regards to immune function. Among the other genes in these windows with high HKA scores are novel candidates for being targets of long-standing balancing selection.

2.3. Non-protein-coding exons

The pattern we observe across the HKA empirical distribution for the percentage of windows overlapping non-protein-coding exons is strikingly different (supplementary fig S5 and supplementary table S4). There is no trend for an increasing percentage of windows overlapping exons as the HKA score decreases. Instead, the percentage of windows overlapping exons is flat along the middle 80% of its distribution, consistent with weaker influence of purifying selection. This may be due to weaker natural

selection in non-protein-coding exons, a lower percentage of windows overlapping conserved elements (as non-protein-coding RNAs tend to be shorter than protein-coding ones), or a combination of both.

2.4. The percentage of protein-coding exons in different bins of the HKA empirical distribution correlates with effective population size

We observe some differences between populations in the patterns above, which may be explained by their differences in effective population size (N_e). The effective population size influences the effectiveness of natural selection (Lanfear et al. 2014). For example, with larger N_e slightly deleterious alleles are more effectively removed from the population through negative selection, and advantageous alleles are more likely to increase in frequency through positive selection. A consequence of negative selection is background selection, the removal of linked neutral variation along with truly deleterious alleles due to linkage. Therefore we expect a greater relative deficit in diversity in and around conserved regions in populations with larger effective population sizes as both positive and purifying selection act to lower diversity.

To test whether the differences observed between populations correlate with their effective population size we used two different estimates of N_e previously calculated from this data set (Prado-Martinez et al. 2013). The first estimate of N_e used is an estimate of long-term N_e calculated based on Watterson's estimator (supplementary table S3 & Prado-Martinez et al. 2013). The second estimate of N_e corresponds to the size of the population since the split with its closest population in the dataset. This was previously estimated using the method of PSMC (Li & Durbin 2011) (supplementary table 5 of Prado-Martinez et al. 2013). We used two different estimates because they measure the effective population size during different evolutionary periods. We were interested in whether either recent or rather long-term N_e shows a stronger correlation with our power to detect the signatures of natural selection and/or with the efficacy of selection. Recent changes in population size can have dramatic effects on levels of genetic diversity, in turn affecting our power to detect localized reduction in diversity. If the differences we observe among populations are largely due to these power issues (higher power to detect local reduction of diversity in populations with a higher overall level of diversity) recent N_e should show the strongest correlation with

the relative reduction in diversity around conserved sequences (protein-coding exons).

The results of a Pearson's correlation test show that the difference between populations in the percentage of windows containing protein-coding exons (E) in a given bin of the HKA empirical distribution significantly positively correlates with their estimated long-term N_e (Figure 2 and supplementary table S3) for the 0.05% and 1-20% bins (0.1% bin marginally non-significant, p-value: 0.07). There is a significant negative correlation between E and estimated long-term N_e for the 55-99% bins. There is a significant positive correlation between E and an estimate of recent N_e for the 0.05-5% bins (Figure 2 and supplementary table S3). The estimate of recent N_e significantly correlates with E in fewer bins than the estimate of long-term N_e (4 compared to 13 bins) and the correlation with recent N_e is only stronger than the correlation with long-term N_e in the bottom 0.05-0.1% bins, suggesting that overall we observe the effects of the relatively long-term evolutionary history of each population, rather than merely differences in power due to the overall level of diversity. Nevertheless, putative differences in accuracy between the two N_e estimates may affect their comparison here.

The correlation notably changes from a positive correlation to a negative one when comparing the regions in the bottom and in the top half of the HKA empirical distribution. We observe a strong positive correlation for several bins in the bottom half of the HKA empirical distribution, which likely reflects the increased effect of negative selection lowering diversity in populations with historically large effective population sizes (possibly combined with our higher power to detect that reduction in samples with higher genetic diversity). This is particularly striking in the case of *P. paniscus*, which has among the lowest N_e of any of the lineages apart from *H. sapiens* and *P.t. verus* (according to the N_e estimate based on Watterson's estimator) and shows no trend for reduced diversity in regions overlapping protein-coding exons. Interestingly, *P. paniscus* showed little effect of purifying selection for loss-of-function variants, whose detection is not dependent on their levels of diversity (Prado-Martinez et al. 2013). On the other hand, populations with historically large effective population sizes have a greater number of exons in regions with low diversity, likely reflecting the actions of purifying and background selection in removing deleterious and linked diversity in these regions respectively.

The strong negative correlation for most of the top half of the HKA empirical distribution also likely reflects less efficient purifying selection and weaker background selection in populations with low effective population sizes, combined with the fact that we may have higher power to detect unusual localized diversity

peaks in populations with lower overall genetic diversity (due to higher N_e).

3. B-scores support the influence of positive and balancing selection in the tails of the HKA empirical distribution.

To further investigate the role of purifying and background selection in driving the different HKA scores observed between regions we tested whether regions with low HKA scores tend to be highly conserved. To do this we calculated the average B score per base of every genomic region for which we have an HKA score. The B score is a measure of the amount of background selection operating on a region, with a lower B score representing a higher level of background selection and being the same for a given position in all populations (McVicker et al. 2009). If purifying selection is shaping patterns of genetic diversity in the genome by removing deleterious variants, then we should observe that as the average HKA score of regions decreases (due to a reduction in diversity) the average B score of these regions also decreases. This general pattern is indeed what we observe (supplementary fig S6 and supplementary table S5) across the HKA empirical distribution if we exclude the tails (between the 5-99% range of the HKA empirical distribution), suggesting that the relative influence of purifying and background selection, which is greatly influenced by the effective population size, may play a role in shaping the relative levels of genetic diversity between the lineages studied.

We observe an upturn in the average B scores of regions in the 5% or lower tail of the HKA empirical distribution in all populations, showing that many of the regions with the lowest HKA scores do not have as low a B score as expected given their extreme reductions in diversity relative to the genome-wide distribution. This might be due to noise, as there are few of these regions relative to larger cutoffs in the HKA distribution. Alternatively, it could be due to strong positive selection acting on these regions, reducing their levels of diversity much more than predicted based on negative selection (B-score) alone. Finally, this pattern could be explained by the relatively large selective sweeps, as diversity is removed in unconserved regions that are close to the selected variants due to hitch-hiking. In any case, the data suggests that we cannot explain well the regions in tails taking into account only purifying selection, supporting the idea that positive selection has contributed significantly to the tails.

We were interested in investigating the effect of N_e in the correlation between a window's HKA

score and its B-score. A Pearson's Correlation test comparing the correlation of the average B-scores per window for different cut-offs in the HKA empirical distribution with the two different estimates of N_e described above shows that the average B-score correlates significantly with N_e from the 1-10% tail for both estimates of N_e (supplementary fig S6 and supplementary table S5). The N_e values from Watterson's estimator show a significant negative correlation with N_e from the 0.05-40% range of the HKA empirical distribution and a significant positive correlation from the 50-99% range. This suggests that effective populations size likely has a significant effect on the ability of the different populations to remove deleterious variants from conserved functional regions.

The 5% tail of the empirical distribution shows the strongest negative correlation with N_e (calculated using Watterson's estimator) with an R of -0.9 and a p -value of 0.0009. This supports the hypothesis that the N_e of a population has a very strong influence on its ability to remove variation in conserved regions through either purifying selection or selective sweeps due to positive selection, although differences in power among the populations may also play a role here.

4. Spatial clustering of windows supports the influence of balancing and positive selection in the top and bottom tails of the HKA empirical distribution.

Neutral evolutionary processes are expected to act uniformly across the genome while selection is considered to be locus-specific. If the windows we observe in the extreme tails of the HKA empirical distribution are the result of drift we would not expect to see increased clustering of these windows in terms of their spatial proximity in the genome relative to the clustering of regions from other parts of the HKA distribution. Therefore any evidence of an increase in spatial clustering of windows in the tails relative to what is observed across the HKA empirical distribution would be indicative of selection acting at specific loci in the genome greater than 30kb in size.

To investigate whether we find any such evidence of spatial clustering we calculated the average distance between all possible pairs of windows in the top and bottom 0.1 % of the HKA empirical distribution respectively for each population. This distance was only calculated between windows on the same chromosome. To determine whether this average distance showed evidence of increased clustered

relative to the neutral expectation we compared it to the average distance between an equal number of regions randomly selected from the remaining 99.8% of the HKA empirical distribution. We observe an increase in spatial clustering in the 0.1% tails of the HKA empirical distribution relative to the central 99.8% for almost all populations (supplementary table S12). This suggests that the extreme tails of the HKA distribution are enriched for regions that contain targets of selective processes. We compared the mean mapping quality and depth of coverage of sites in these regions to randomly sampled regions from the genome-wide distribution. We find no evidence that regions in the extreme tails of the HKA distribution are unusual in these regards (see Supplementary Materials HKA 1.3 and 1.4).

The only exception to this pattern is the 99.9% tail of the HKA empirical distribution in *P.t. schweinfurthii*, where the windows do not show an excess of spatial clustering relative to the central 99.8%. However closer inspection of the location of the windows in the 99.9% tail shows that there is a strong pattern of clustering at the chromosomal level, with 42 windows on chromosome 6. The chromosome with the 2nd largest number of windows is chromosome 9, which has only 16 windows in the 99% tail. The high number of windows on chromosome 6 in the 99.9% tail of the HKA empirical distribution is a feature shared across populations and is likely to primarily reflect long term balancing selection acting on the MHC region. As *P.t. schweinfurthii* has several chromosomes with only two or three windows in the 99.9% tail of the HKA empirical distribution it is possible that the large distance between these windows masks the signal of spatial clustering of windows under balancing selection in the MHC region and other potential cases of balancing selection.

The increased spatial clustering of windows in the 0.1% tail of the HKA empirical distribution relative to the general distribution is observed across all populations (supplementary table S12). This may be due to positive selection causing selective sweeps, which result in the loss of diversity in a wide region flanking the selected variant. This observation may explain the reduction in exonic regions in the extreme bottom tail of the HKA empirical distribution we report above (see Supplementary Materials HKA 2.4), if genic targets of selection are flanked by selective sweeps with similarly low levels of diversity and high levels of divergence. Alternatively this may be due to the strongest targets of selection being non-genic.

To investigate if these results were related to differences in N_e between the populations we ran a Pearson's Correlation test with two different measures of N_e described above (supplementary table S12). Although we observe no significant correlations with either estimate of N_e , the amount of spatial clustering

in the bottom 0.1% of the HKA empirical distribution approaches significance with a p-value of 0.06 and an R score of -0.6 when using the long-term N_e values based on Watterson's estimator. This suggests that N_e may be influencing the strength and/or number of selective sweeps that occur in a population, with more and/or stronger selective sweeps occurring in populations with larger N_e . These results suggest that among the *Hominidae* selective sweeps may be more prevalent in populations with larger long term N_e , although differences in power between species with different levels of diversity could also play a role.

5. Correlation of results with Tajima's D

To search for additional evidence that the candidate regions we identified as evolving under positive and balancing selection were true positives we calculated Tajima's D, another commonly used test for detecting positive and balancing selection, genome-wide for each species and performed permutation testing to explore Tajima's D signatures in the 1% extreme of the FWH (one-tail, see Supplementary Materials Section 3) and HKA (two-tails) empirical distributions. The results (supplementary table S107) show that windows in the tails of the FWH and HKA distribution are highly enriched for extreme values of Tajima's D (highly significant in 25 out of 27 comparisons). The Tajima's D test thus identifies similar outlier genes to the FWH and HKA tests and demonstrates that these results are broadly consistent with another test for selection.

6. Candidate selected regions and genes

For each population we produced two separate tables of the 200 protein coding genes with the strongest signal of positive or balancing selection as determined by the rank of the windows they overlap in the HKA empirical distribution. Genes in windows with the lowest HKA score in the empirical distribution are candidates for being under positive selection while genes in windows with the highest HKA scores are candidates for being under balancing selection. These tables can be found in supplementary table S6. For each gene we report the rank of the window it overlapped, its chromosomal location, the HKA score of the window, the p-value of the window based on its rank in the HKA empirical distribution, and further

information about the gene including its Ensembl ID and gene name, the gene type (e.g. protein-coding, lincRNA etc) and how many base pairs of exons overlapped the window.

7. Functional annotation in candidate regions

We sought to identify whether the regions with signatures of positive selection in the HKA test are enriched for functional elements, such as protein coding genes. To identify any potential signals of functional enrichment among the regions with signatures of positive selection in the HKA test we calculated the percentage of windows in the 0.1% tail containing either functional annotation, protein-coding exons or non-protein-coding exons. Windows were annotated using GENCODE hg18 annotation (Harrow et al. 2012). Across all lineages values ranged from 62-90% for windows containing any type of functional annotation, 50-85% for protein coding exons (a subset of functional annotation) and 15-24% for non-protein coding exons.

To test whether these values were greater than expected for each lineage we performed random sampling of an equal number of windows as found in the 0.1% tail from across the genome-wide distribution of 30kb windows. From these regions we similarly calculated the percentage of regions containing functional annotation. This process was repeated 100 times for each lineage. The results of these random permutations were compared to the results in the 0.1% tail (supplementary figs S7-S15). We find that the proportion of candidate windows with signatures of positive selection that overlap functionally annotated elements or protein coding exons is significantly greater than the proportion that these annotation categories represent in the genome for the majority of lineages except *P. paniscus*, *P.t. verus* and *P. pygmaeus*. In contrast, no lineage showed an enrichment of non-protein coding genes in the 0.1% compared to sub sampling from across the genome. These results suggest that the candidate regions do not simply reflect the tail of a neutral distribution and provides clues as to the targets of adaptive evolution.

8. Length bias of GO categories

As the HKA and FWH tests are performed using genomic windows there is a potential for variation in gene length to bias the results of category enrichment. Assuming a null hypothesis of no selection acting on the

genome, then the genes in the extreme tail of the distribution may be overrepresented by long genes, because these cover more genomic windows and are therefore more likely to appear in the tail by chance. Any biological categories that are enriched for long genes may therefore show signals of significant enrichment due to this length bias. This potential effect of gene length on enrichment tests is rarely accounted for in window-based selection analyses.

To investigate if our results are influenced by this potential source of bias, we calculated, for each lineage, the mean length of the 200 genes in the bottom tail of the HKA empirical distribution. We compared this to the mean gene length of 200 genes in a random sample of genomic windows, 100 times. We observe no evidence of genes in the bottom tail of the HKA distribution being longer than the genomic mean based on the mean of the 100 means generated by this sub sampling procedure.

We repeated the process for the top tail of the HKA empirical distribution (candidate genes for balancing selection). For most lineages the mean gene length falls within the distribution of means obtained from random sampling. However for *P. paniscus* the mean length of genes in the top tail falls above the distribution of means obtained by random sampling. This trend may be partially driven by large intronic regions in these genes, where the absence of purifying selection permits diversity to accumulate.

We additionally explored the potential influence of length bias on our gene category enrichment test results by randomly sampling 200 windows from our genome-wide distribution using the process described above, 100 times. This process was repeated across lineages. For each pseudotest set we ran the GO enrichment analysis using FUNC (a software for biological enrichment analyses that can be run at the command line and is therefore capable of performing this number of tests) (Prüfer et al. 2007). The most frequently enriched GO category in these re-samplings is the 'biological process' category 'homophilic cell adhesion', which appears as significantly enriched in 6 out of the 100 re-samplings in *P. abelii* (supplementary tables S7-S11). Most other categories that appear as significantly enriched appear only once in a given population. In our actual results the GO category 'dendrite' appears as significantly enriched in three lineages (supplementary tables S16-S29. In the random sampling in two lineages this category is significantly enriched by chance in one of the 100 random sampling procedures and in one population it is significantly enriched in three of the 100 cases. The probability of this occurring by chance in all three lineages where we observe significant enrichment is thus extremely low. Therefore, these results suggest that gene length bias is unlikely to affect our enrichment of biological categories results (supplementary tables

S7-S11).

References

- Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical research* **70**: 155–174.
- Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research* **22**: 1760–1774.
- Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–159.
- Hughes AL, Yeager M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Annual review of genetics* **32**: 415–435.
- Kaplan NL, Darden T, Hudson RR. 1988. The coalescent process in models with selection. *Genetics* **120**: 819–829. <http://www.genetics.org/content/120/3/819.short>. H, Eyre-Walker A. 2014. Population size and the rate of evolution. *Trends in ecology & evolution* **29**: 33–41.
- Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favorable gene. *Genetical Research* **23**: 23–35.
- McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics* **5**.
- Prüfer K, Muetzel B, Do H-H, Weiss G, Khaitovich P, Rahm E, Pääbo S, Lachmann M, Enard W. 2007. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC bioinformatics* **8**: 41.
- Zhai W, Nielsen R, Slatkin M. 2009. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Molecular Biology and Evolution* **26**: 273–283.

Section 3: FWH test

Hafid Laayouni, Marc Pybus, Ferran Casals, Jaume Bertranpetit

1. Fay Wu's H statistic calculation:

The Fay and Wu H statistic (FWH; Fay and Wu, 2000) measures departures from neutrality reflected in the difference between derived segregating sites at high-frequency and intermediate-frequency alleles. This statistic is especially robust to population demographic history, making it an improvement over Tajima's D in this regard (Tajima, 1989). FWH takes advantage of ancestral information and is implemented in an overlapping window approach. The algorithm is run in windows of 30Kb with an offset of 15Kb. We validated the algorithm using simulations under neutral and selective scenarios with human demographic parameters (supplementary fig S19).

Windows with less than 300 callable sites were removed from the analysis to increase the quality of the data. Finally, an empirical p-value is calculated for each window for which we have a score thus generating a genome-wide ranking of scores.

2. Results

The Fay Wu's H algorithm was run on each lineage. In order to obtain a list of genes under positive selection for each lineage, windows were annotated following the same procedure as for the HKA pipeline (see Supplementary Materials HKA).

Using the empirical p-value of the FWH score, we first focus on the 1% extreme distribution of windows analyzed (approx. 1600 windows). Once annotated, for most lineages approximately 45% of these windows correspond to protein coding genes, 12% are lincRNA, 8% are pseudogene, 15% belong to other functional elements while 20% are not annotated as functional elements.

2.1 Genes with extreme footprints of selection

The supplementary table S74 lists the top candidate gene targets of positive selection for the lineages

analyzed. All genes are provided with a rank value for the FWH statistic. Ranks correspond to the empirical distribution of windows in the analyzed genome. Many genes involved in immune response appear among the top of genes putatively having evolved under positive selection (which we refer to as outliers with reference to the empirical distribution). Some of these genes also belong to the 1% extreme tail of many lineages analyzed. For example, *FER* (cytokine-mediated signaling pathway) is an outlier in *G.g. gorilla*, *P. abelii* and *P. pygmaeus* and in *P.t. ellioti*. *STAB2* (defense response to bacterium) is outlier in *P. paniscus*, *P.t. ellioti* and *P. abelii*. *HLA-DMA* (immunoglobulin mediated immune response) is outlier in *P.t. verus* and *P.t. schweinfurthii*. All the other genes involved in immune response appearing in the 200 top genes are outliers in two lineages or are specific to one lineage, a finding expected if there is stratification in the pathogenic environment among groups.

Many genes involved in neurobiological processes appear among the top genes putatively evolving under positive selection. *KCNIP4* (neuronal cell body and signal transduction) belong to this extreme 1% in *P. paniscus*, *P. t. troglodytes*, *P.t. schweinfurthii*, and *P.t. verus*. *ITGA8* (nervous system development; memory) belong to the 1% extreme tail of the distribution in *P. paniscus*, *P.t. verus*, *P.t. schweinfurthii* and *P.t. ellioti*. *FAM169B* (neuronal cell body and signal transduction) is an outlier in *P.t. troglodytes*. *GPR98* (neurological system process; sensory perception of sound) is an outlier in *P.t. schweinfurthii*. *NELL1* (nervous system development) is an outlier in *P. paniscus*. *FGF14* is also involved in the development of nervous system and is an outlier in *P. paniscus*, *G.g. gorilla* and *P.t. verus*. *NRG3* (member of the neuregulin gene family and implicated in susceptibility schizophrenia and schizoaffective disorder.) is an outlier in *P. abelii* and *P. pygmaeus*, *P. t. ellioti*, *P.t. schweinfurthii* and *P.t. troglodytes*. *NRXN3*, a gene encoding a member of a family of proteins that function in the nervous system as receptors and cell adhesion molecules shows up as extreme in three *Pan troglodytes* subspecies (*P.t. schweinfurthii*, *P.t. troglodytes*, *P.t. ellioti*) and in *P. pygmaeus*.

Interestingly, many genes involved in reproductive processes show up among the targets of positive selection; *LGR4* is involved in the development of male genitalia and is an outlier in the *P.t. troglodytes* lineages. *BARD1* is involved in spermatogenesis and is an outlier in *P.t. schweinfurthii* and *P.t. ellioti*. *HSD17B4* is involved in androgen and estrogen metabolic processes and is an outlier in *P. paniscus* and *P. abelii*. *RAD23B* is involved in spermatogenesis and is an outlier in *P. abelii*. *SPAMI* is involved in fusion of

sperm to egg plasma membrane and is an outlier in *G.g. gorilla*. *IQCJ-SCHIP1* is involved in female gonadal development and is an us outlier in *G.g. gorilla*, *P. paniscus*, *P. abelii*, and *P.t. schweinfurthii*. Worth noting is that *PCDH15*, a gene involved in adult walking behavior and visual perception, shows up as an outlier in *P.t. schweinfurthii* and *P.t. ellioti* and *P. paniscus*.

The 200 coding genes in the 1% tail of the FWH distribution, containing putative signals of positive selection, show considerable overlap across lineages (supplementary table S73). The average pair-wise intersection is 7.4% (ranging from 3% between *G.g. gorilla* and *P. abelii*, and between *P.t. troglodytes* and *P. pygmaeus* to 25% between *P.t. troglodytes* and *P.t. schweinfurthii*). This suggests that there are multiple cases of shared genic targets of positive selection between lineages.

References

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics*, 155, 1405–1413.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 2005 Nov;15(11):1576-83.

Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–95.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 2010;11(2):R14.

Section 4: MK test

Gabriel Santpere, Arcadi Navarro

1. Methods

1.1. Selection scan

We carry out a scan of natural selection in each of the studied lineages. To do so we focus on protein-coding genes and use the classical McDonald-Kreitman (MK) test (McDonald & Kreitman, 1991). This test compares the variation accumulated in a given species to the divergence between this and other species at two site classes (e-g. at synonymous versus non-synonymous sites). This first version of the test assumes that all non-synonymous changes are neutral, strongly advantageous or strongly deleterious. Considered this way, strongly deleterious mutation would have been eliminated by selection while strongly advantageous ones would have been fixed or contribute very little to polymorphism. Thus, existing diversity at non-synonymous sites is assumed to be mostly neutral. Another assumption of the test is that evolution occurred in a diploid panmictic scenario with stationary size. The MK test has been shown to be more powered to detect negative rather than positive selection using simulated *H. sapiens-P. troglodytes* genetic data (Zhai, Nielsen, & Slatkin, 2009).

1.2. Samples

For the MK test scan, we considered only species with at least five individuals (in the case of *P.t.verus*, we included the admixed Donald). We excluded individuals based on coverage to increase the proportion of the genome that could be evaluated (supplementary table S93 and Tables S1-S2).

1.3. Protein-coding sequences for selection scan

We extract coordinates for the coding part of all coding autosomal transcripts annotated in the RefSeq hg18, considering only unique identifiers (n=31,703, covering ~32Mb of genome).

We then intersect these CDS coordinates with the callable portion of the unique part of the genome that considers a position if all individuals used in the analyses show a minimum coverage of 5x. This leaves ~15.1Mb of coding sequence (collapsing overlapping transcripts).

For each species and coding sequence we count all present SNVs and substitutions (supplementary table S94). We require both SNVs and substitutions to have appeared after the split with the most recent common ancestral node with *H. sapiens* (in the case of *H. sapiens*, we take the ancestral node with *Pan*). We oriented variants using the reconstruction of variant ancestralities from (Prado-Martinez et al., 2013). For the *H. sapiens*, *Pan* and *Gorilla* genus we only consider variants that present a monomorphic ancestral node. For *Pongo*, we get only those variants that are different from a monomorphic ancestral node between *Gorilla* and *H. sapiens*, and from *Macaca mulatta*. Missing data was allowed in the case of counting a position with a SNP but no missing data was allowed when counting substitutions.

1.4. Variant annotation and MK test

We use the ANNOVAR software (<http://www.openbioinformatics.org/annovar/>) to annotate the synonymous (S)/non-synonymous (NS) effect of all SNPs and substitutions occurring in a transcript. Positions with more than two alleles were excluded.

We then count how many S and NS polymorphisms and substitutions we observe in each transcript. We then construct a contingency table and test for association with a one-tailed Fisher exact test. The test is performed only when a transcript shows three or more S changes (Ps+Ds) and three or more NS changes (Pn+Dn). For each transcript we obtain a p-value and an odds ratio (OR). For OR calculations, we added 0.5 to all cells if one of the cells was zero. In supplementary table S95 we present all transcripts evaluated for each species.

1.5. Estimates of rate of adaptive substitutions in the different primate species

We estimate the proportion of adaptive substitutions (α) and their rate ($\omega\alpha$) in all studied species, using an extension of the MK test (Eyre-Walker & Keightley, 2009) as implemented in the DFE-alpha software. This method compares the site frequency spectrum from neutral and selected sites and infers a distribution of fitness effects of new deleterious mutations by maximum likelihood, and models recent demographic changes, to emit an estimation of α and $\omega\alpha$. This more sophisticated method has limitations in the number of sites evaluated to give proper estimates, and makes it unsuitable to study individual genes. To estimate the rate of protein adaptation in the different primate species we used a concatenated data-set of 3859 orthologous genes non-overlapping with any other described transcribed genomic elements in the human genome according to Refseq. 4fold synonymous sites were used as a neutral reference to estimate the proportion and rate of adaptive substitutions in 0fold non-synonymous sites using the DFE-alpha software. Unfolded site frequency spectra for each primate species were obtained using ancestral states in the closer node splitting each species with *H. sapiens*, as described for the classical MK test scan. Correlation between N_e and various DFE-Alpha estimates was studied while attending to phylogenetic non-independence of the traits in the tree of the species analyzed using BayesTraitsV2, setting the methods to employ random-walk and maximum likelihood. Significance was assessed by comparing a model with a free correlation with another model with correlation fixed to 0, by means of a likelihood ratio statistic.

2. Results

2.1. Significant genes in MK scans

We tested for genes significant in the MK test at both tails indicating genes especially constrained (under strong purifying selection) and genes putatively under positive selection. In supplementary table S96 the number of genes significant in the MK test with $p < 0.05$ are shown. The MK test is more powered to detect instances of strong purifying selection (Zhai et al., 2009) as observed by the increased number of significant transcripts obtained. P-log columns indicate transcripts that are significant and at the same time possess an extreme ($p < 0.05$) OR value. The complete list of significant transcripts can be found in supplementary table S97.

We compared our results with a previous MK genome-wide scan (Bustamante et al., 2005). We obtained from this study a list of genes described as positively selected (115) or negatively selected (215) between *H. sapiens* and *P. troglodytes*. Fifty positively selected genes in Bustamante et al. (2005) were also evaluated in our study and none of them gave a significant signal for positive selection with a p-value < 0.05. Of the 215 genes determined to be under strong purifying selection we had power to evaluate 147, and found four genes (P2RX7, TIAM1, COL12A1, CNGB3) that were also significant in our MK test. For the overlapping genes in both studies we performed a density plot of the logOR obtained for these genes in our analysis (supplementary fig S16). Clearly, genes with described strong purifying selection in Bustamante et al., 2005 are skewed to the left and separated from the overall distribution of logOR in our study. The distribution of logOR of genes with described positive selection appear bimodal with one peak skewed to the right and one other peak overlapping the average of the total gene-set. In general, the overlap between both studies is higher regarding purifying selection.

Several significant genes are related to nervous system function (supplementary table S98) and development. In particular *MCPHI*, *CASC5*, *PHGDH*, *FTO* and *NBN*, when mutated in *H. sapiens*, carry associated changes in brain size. Variants in other genes, such as *SETX*, *MTPAP*, *RNF213* or *VCAN* are related to neurodegenerative processes in *H. sapiens*. Interestingly, *SETX*, a gene involved in spinocerebellar ataxia and amyotrophic lateral sclerosis, has been also described to be a target of recent positive selection in CEU (Grossman et al., 2013). Another well-represented group of significant genes are related to the immune system (supplementary table S99). Strikingly, we find several genes that help in the defense against viruses, such as *ZC3HAV1*, *HIVEP1*, *MX1*. Finally, we obtained a list of other significant genes related to several other human disorders (supplementary table S100).

Lifespan and attributes of the senescence process are important divergent features among primates. Longevity in humans is notably enhanced compared to non-human primates even if they are in captivity; the oldest living human was 122 years old compared to the record of 74 years for chimpanzees (de Magalhães & Church, 2007). However, the genetic bases of these differences are yet to be revealed. Many comparative studies on ageing have been performed on Rhesus monkeys but much less is known about the ageing and cognitive decline in great apes, with chimpanzees being the most studied. We have identified ageing-related genes with signatures of positive selection particularly in *Pongo*, which present a maximum lifespan in captivity of around 59 years. For example we detected one gene related to ageing, *WRN*, with evidence of

positive selection in *P. pygmaeus* using the MK test. Mutations in *WRN* in humans can cause Werner syndrome which is a dramatic progeroid syndrome, i.e. causes premature ageing (Goto, 1997). Interestingly, previous studies reported possible adaptive acceleration in *WRN* since humans and chimpanzee diverged (Clark et al., 2003; de Magalhães & Church, 2007). We also found *NBN* under positive selection in *P. pygmaeus*, a gene also considered to be involved in progeroid syndromes in humans. Mutations in *NBN* cause a chromosomal instability syndrome called Nijmegen breakage syndrome (Martin & Oshima, 2000), that is accompanied by features of senescence. *P. abelii* shows also signals of positive selection in *ERCC5*. ERCC proteins play a role in DNA repair and have been linked to senescence. Mutations in *ERCC6* and *ERCC8* have been reported as causes of premature ageing in humans (de Magalhães & Church, 2007). Finally, *PPM1D* shows evidences of positive selection in *P. t. ellioti*, which has been shown to reduce longevity in *PPM1D*-null mice, specifically in males (Nannenga et al., 2006).

2.2. Functional enrichment analysis

We performed a functional enrichment analysis with WebGestalt (<http://bioinfo.vanderbilt.edu/webgestalt/>) to test for gene set enrichment in functional categories in *non-slimmed* Gene Ontology, KEGG pathways, Pathway Commons, Wikipathways, Disease (PharmGKB and GLAD4U) and PheWAS associated. In this case, according to a Bonferroni correction adjusted p-value < 0.05 and requiring at least two genes to be present in one category, we obtained several enriched categories (supplementary table S101). For instance, 'pancreatic function' appeared enriched in positively selected genes in *P.t. ellioti*, *P.t. schweinfurthii* and *P. pygmaeus*. *P. paniscus* selected genes showed an enrichment in 'potassium voltage-gated channel activity' and also in diseases of the nervous system such as 'Amyotrophic Lateral Sclerosis'. Gorilla selected genes are enriched in the 'glycoprotein metabolic process' and 'glycosaminoglycan binding' categories, and also some of these same genes drove significance to the *cartilage disease* category.

Since the small number of significant transcripts obtained precludes a properly powered functional enrichment analyses we concatenated all genes belonging to each PANTHER pathway (Thomas et al. 2003) and performed the MK test. In supplementary table S102 the p-values obtained for each of these pathways are shown. We observed that many pathways were under strong negative selection in many great ape species

at the same time ($p\text{-value} \leq 0.05$). In particular, 'Alzheimer's disease-presenilin pathway', 'Integrin signaling pathway' and 'Wnt signaling pathway' appeared significantly constrained in all lineages. The Integrin pathway includes the components involved in the downstream events triggered by the interaction of integrins with elements of the extracellular matrix, such as actin related genes and MAPKs. Actin and MAPK related KEGG pathways have also been reported to be enriched in genes under purifying selection between human and chimpanzees and between mouse and rats in previous studies (Serra et al. 2011). The Wnt pathway has been related to many important biological processes and may have a universal role in configuring the primary axis of animals (Nusse & Varmus, 2012). The Alzheimer's disease (AD) pathway includes genes involved in this human neurodegenerative disease. In agreement with our results, crucial AD genes such as *APP* and *MAPT* were reported to be conserved between human and chimpanzees (Hamilton, 2004; Holzer, Craxton, Jakes, Arendt, & Goedert, 2004; Rosen et al., 2008). This is an interesting finding because although the main neuropathological hallmarks of AD (i.e. A β and hyper-phosphorylated tau deposition) have also been observed in the ageing brain of chimpanzees (Gearing, Rebeck, Hyman, Tigges, & Mirra, 1994; Rosen et al., 2008), the complete AD clinical and neuropathology seems to be presented by humans only.

A few pathways showed a nominally significant p -value for positive selection: i.e. 'Plasminogen activating cascade' in *P.t. verus* and *P.t. schweinfurthii*, 'Axon guidance mediated by semaphorins' also in *P.t. schweinfurthii*, 'Glutamine glutamate conversion' in *P. paniscus*, 'Serine glycine biosynthesis' in *G.g. gorilla*, and finally the 'Thyrotropin-releasing hormone receptor signaling pathway', 'Formyltetrahydroformate biosynthesis' and the 'Alpha adrenergic receptor signaling pathway' in *P. abelii*. But only one pathways, 'Plasminogen activating cascade', appeared in more than one lineage: *P.t. verus* and *P.t. schweinfurthii*.

2.3. Correlation between rate of adaptive substitutions (ω) and effective population size (N_e).

Proportion of adaptive substitutions (α) and rate of adaptive substitutions ($\omega(\alpha)$) were estimated using DFE-alpha by combining its estimated distribution of fitness effects (out of the species SNP data) and each species derived substitutions, considering modeled recent demography/sweeps effects. The supplementary table S103 shows the estimates of α and jointly with estimates of synonymous and non-synonymous polymorphism (θ_s and θ_n) and the evolutionary rate (omega, ω) in non-synonymous 0fold sites. The values for N_e calculated from Watterson's θ , from (Prado-Martinez et al., 2013) are also indicated. We

obtained in general low values of both α and $\omega(\alpha)$, in agreement with previous estimates (Eyre-Walker, 2006; McManus et al., 2015). *P.t. schweinfurthii* presented the highest values probably as the result of including two of the six individuals (Vincent and Andromeda) with high levels of inbreeding.

According to the nearly neutral theory, selection efficiency depends on the effective population size, because the fate of a mutation is determined by the product NeS (Ohta, 1976; Tomoko Ohta, 2002)(Ohta 1976,2002; Lynch and Connery 2003). We check this theory correlating the estimates of rate of adaptive substitutions with the estimates of effective population size for each primate species. Correlations were calculated while controlling for the phylogenetic non-independence using the generalized least square approach (Table S105) implemented in BayesTraitsV2 and using the random walk/maximum likelihood method. We obtained a measure of significance by comparing the likelihood of the model with free correlation value with a model with a correlation fixed to 0. The correlation is positive for alpha and omega(alpha) but non-significant. The observed rate of adaptive evolution in *P. pygmaeus* was poorly supported by bootstrap and fell close the 95% quantile; removing them together with *P.t. schweinfurthii*, that included inbreed individuals from a different geographical region and also excluding humans, that presented very different Ne between African and Non-African individuals as in (Prado-Martinez et al., 2013), we obtained a significant correlation between $\omega(\alpha)$ and Ne (supplementary fig S16(C)). This all suggests that primate species with higher effective population size have a higher rate of adaptive evolution in proteins.

Estimated DFE for new mutations at 0-fold sites (supplementary table S104 and supplementary fig S18) shows that most mutations are deleterious at this class of sites with little variation among species (sites with predicted $NeS > 10$ account for more than 65% in all primates). Additionally, we also consistently found a marginally significant negative correlation between proportions of neutral or nearly neutral sites with a long term Ne , indicating that the deleterious effect of mutations is greater in populations with larger Ne , illustrating a more efficient action of selection in the latter.

References

- Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., ... Clark, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature*, *437*(7062), 1153–7. doi:10.1038/nature04240
- Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. a, ... Cargill, M. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science (New York, N.Y.)*, *302*(5652), 1960–3. doi:10.1126/science.1088821
- De Magalhães, J. P., & Church, G. M. (2007). Analyses of human-chimpanzee orthologous gene pairs to explore evolutionary hypotheses of ageing. *Mechanisms of Ageing and Development*, *128*, 355–364. doi:10.1016/j.mad.2007.03.004
- Eyre-Walker, A. (2006). The genomic rate of adaptive evolution. *Trends in Ecology & Evolution*, *21*(10), 569–75. doi:10.1016/j.tree.2006.06.015
- Gearing, M., Rebeck, G. W., Hyman, B. T., Tigges, J., & Mirra, S. S. (1994). Neuropathology and apolipoprotein E profile of aged chimpanzees: implications for Alzheimer disease. *Proc Natl Acad Sci U S A*, *91*(20), 9382–9386.
- Goto, M. (1997). Hierarchical deterioration of body systems in Werner's syndrome: implications for normal ageing. *Mechanisms of Ageing and Development*, *98*(3), 239–54. Retrieved from _____
- Hamilton, B. A. (2004). alpha-Synuclein A53T substitution associated with Parkinson disease also marks the divergence of Old World and New World primates. *Genomics*, *83*(4), 739–742.
- Holzer, M., Craxton, M., Jakes, R., Arendt, T., & Goedert, M. (2004). Tau gene (MAPT) sequence variation among primates. *Gene*, *341*, 313–322.

Martin, G. M., & Oshima, J. (2000). Lessons from human progeroid syndromes. *Nature*, *408*(6809), 263–6.

doi:10.1038/35041705

McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, Ryder OA, Ape Genome Project G, Kidd JM, Wall JD, et al. 2015. Inference of gorilla demographic and selective history from whole-genome sequence data. *Molecular biology and evolution* **32**: 600–12.

<http://mbe.oxfordjournals.org/content/32/3/600.short>.

Nannenga, B., Lu, X., Dumble, M., Van Maanen, M., Nguyen, T.-A., Sutton, R., ... Donehower, L. A.

(2006). Augmented cancer resistance and DNA damage response phenotypes in PPM1D null mice.

Molecular Carcinogenesis, *45*(8), 594–604. doi:10.1002/mc.20195

Nusse, R., & Varmus, H. (2012). Three decades of Wnts: a personal perspective on how a scientific field developed. *The EMBO Journal*, *31*(12), 2670–84. doi:10.1038/emboj.2012.146

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., ... Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–475.

Rosen, R. F., Farberg, A. S., Gearing, M., Dooyema, J., Long, P. M., Anderson, D. C., ... Walker, L. C.

(2008). Tauopathy with paired helical filaments in an aged chimpanzee. *J Comp Neurol*, *509*(3), 259–270.

Serra, F., Arbiza, L., Dopazo, J., & Dopazo, H. (2011). Natural selection on functional modules, a genome-wide analysis. *PLoS Comput Biol*, *7*(3), e1001093.

Sidow, A. (1992). Diversification of the Wnt gene family on the ancestral lineage of vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(11), 5098–102. Retrieved

from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?>

artid=49236&tool=pmcentrez&rendertype=abstract

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome research* **13**: 2129–41.

Section 5: ELS test

Alexander Cagan, Christoph Theunert, Kay Prüfer, Aida M. Andrés

1. Extended Lineage Sorting (ELS) Test and Motivation

We aim to identify lineage-specific targets of natural selection by identifying regions of the genome where variation falls external to an outgroup species and which are longer than expected under neutrality. For a detailed description of the Extended Lineage Sorting test see Supplementary Information 7 in Prüfer et al. (2012) and also SOM 13 in Green et al. (2010) where it was first used. Briefly, strong positive selection on a new mutation is expected to cause a selective sweep at that locus. This pattern is produced by the haplotype carrying the selected allele rising to fixation, removing neutral diversity in the region. The effect is equivalent to a bottleneck of one individual around this locus, as all lineages in the population will coalesce at the time of the sweep. Over time variation in this region will recover due to new mutations and recombination. The new variation will be unique to this lineage. Thus, when comparing diversity in this region in relation to an outgroup, the variation in the outgroup will always fall outside the variation in the ingroup (external). When two sister species with a divergence time that is recent enough for the majority of coalescent events to occur in the common ancestor (internal) are compared, the presence of large external regions can thus be used as a signal for detecting positive selection. The size of external regions is a product of the selection coefficient and the recombination rate. The stronger the selection and the lower the recombination rate, the longer we expect the external regions to be. By controlling for variation in recombination rate it should be possible to identify regions based on the strength of selection they have experienced since their divergence with the species used as an outgroup. It is worth noting that purifying selection is also expected to result in shallow trees and external regions. Nevertheless, under purifying selection the external signal is not expected to extend to the same extent as under positive selection, and thus we expect the longest external regions in the genome to be highly enriched for targets of positive selection.

This method allows for the detection of regions that experienced positive selection at a time depth not usually reached by other tests based on the departure of present-day intra-specific diversity from neutral

expectations, such as those based on haplotype length (where the signal of selection decays once the sweep has reached fixation) or the site frequency spectrum (where the signal fades as the locus reaches equilibrium). Therefore this method has the potential to detect lineage-specific positive selection with a greater time depth than many other selection tests which rely on selection being ongoing or very recent. In some species, this includes selection around the time of speciation, which could have played an important role in creating species-specific adaptive traits.

1.1. Pre-test Filtering

We applied a series of pre-test filtering steps to make the data as comparable as possible within and between populations. Variability in the number of individuals with genotype calls across sites results in unequal power to detect selection (for general pre-test filtering see Methods). For each comparison the dataset is further filtered to only include sites where at least one of the lineages is polymorphic, as these are the most informative sites for the test.

1.2. Performing the ELS Test

The ELS test requires two lineages as input. The test is most powerful when the divergence time of these lineages is recent enough that the majority of loci have a coalescence time in the common ancestor of the two lineages (internal) but where there has been sufficient time for diversity to recover after lineage-specific selection (external).

For each comparison, neutral coalescent simulations were performed using the program *ms* to infer the expected fraction of the genome for which the tested population falls inside the variation of the outgroup (internal) compared to outside this variation (external) (Hudson 2002). Simulations were performed using a uniform recombination rate. The simulations require demographic parameters of the lineages being compared, such as estimates of current effective population size for the test and outgroup lineages, mutation rate and generation time (parameters taken from Prado-Martinez et al., 2013; supplementary table S66). The results of the simulations were compared to the observed fraction of derived sites in the data. The simulations were generally found to closely match the observed fraction of derived sites (supplementary fig

S20). The simulations were run for the following comparisons: *P. troglodytes*-*P. paniscus* (outgroup), *P. paniscus*-*P. troglodytes* (outgroup), *G.g. gorilla*-*G.b. graueri* (outgroup), *G.b. graueri*-*G.g. gorilla* (outgroup), *P. abelii*-*P. pygmaeus* (outgroup), *P. pygmaeus*-*P. abelii* (outgroup). We also looked at the *P. troglodytes* sub-species *P.t. ellioti*-*P.t. schweinfurthii* (outgroup), *P.t. schweinfurthii*-*P.t. ellioti* (outgroup), *P.t. troglodytes*-*P.t. ellioti* (outgroup), *P.t. verus*-*P.t. ellioti* (outgroup). For comparisons with *G.g. gorilla* we used *G.b. graueri* although due to the low number of individuals (three) they were excluded from other analyses.

The simulations provided information regarding which species comparisons are adequate for this test, and which ones would be underpowered due to too high a percentage of their genome falling external to the outgroup under neutrality (supplementary table S67). For example, both *P. paniscus* and *P. pygmaeus* had over 75% of their genome falling external to their outgroup under neutral simulations, and thus the test was not run for these species. We also excluded *G. b. graueri* because of our inability to find adequate demographic parameters (the frequencies of derived to ancestral alleles in simulations did not closely match the data, suggesting that the test would produce inaccurate results, supplementary fig S11).

1.3. Hidden Markov Model

A hidden Markov model (HMM) is used to assign all SNP positions with a hidden state of *internal* or *external*. For a full account of the calculation of the emission and transition probabilities see Prüfer et al. (2012). The HMM uses all available individuals from the test lineage and one individual from the outgroup lineage. Due to the presence of multiple individuals in the outgroup populations, the HMM was run repeatedly on the test lineage, using a different individual from the outgroup lineage each time.

After running the HMM, the output is filtered to remove SNPs with a posterior probability < 0.8 for being internal or external. Neighboring SNPs that have been classified as external are merged to form external regions along the genome. Larger regions should be indicative of stronger positive selection. However, region size may also be influenced by recombination rate and low SNP density. To correct for this, regions are scored as a function of their size, the local recombination rate, and physical distance between adjacent SNPs. Specifically, each region is re-scored by calculating the 1Mb average human recombination rate (rr) for the midpoint between any two adjacent SNPs in a region. The recombination map is taken from

Kong et al (2002). Each pair of adjacent SNPs is assigned a value of $1000/rr$ if their physical distance exceeds this value. Otherwise the physical distance is given in base pairs. These values are then summed over all pairs of SNPs and multiplied by the average recombination rate of the entire region. This provides a SNP-corrected value of genetic distance per region.

For each lineage the results of the multiple runs of the HMM (each run obtained using a different outgroup individual) are combined to eliminate the effect of using a single outgroup individual, and thus reduce the number of false positives. This is done as follows: For each run of the HMM, the results are ranked by score as described above. The top 5% of external regions from each run are selected and are given a rank based on their score. Regions are only retained if they appear in the top 5% in all runs (thus, regardless of the outgroup individual used). To refine the detection of the signal, the regions are then trimmed to only the part of the region present in all runs. This list of regions is then ranked based on the cumulative total of the rank score from the multiple runs.

2. Results

The number of unusually long external regions in the 5% extreme of the score distribution varies for each species tested (supplementary tables S68-S70). We detect 27 external regions candidates to contain targets of selection in *P. troglodytes*, 11 regions in *G.g. gorilla*, and 26 regions in *P. abelii*. These regions contain 15 genes in *P. troglodytes*, 7 genes in *Gorilla* and 27 genes in *P. abelii*. The average external region size is 30,723bp for *P. troglodytes*, 29,875bp for *Gorilla* and 65,451bp for *P. abelii*. Many of these regions do not contain genes, suggesting that selection has been acting on both coding and non-coding variation.

References

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neandertal genome. *Science (New York, NY)* **328**: 710–722.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)* **18**: 337–338.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. 2002. A high-resolution recombination map of the human genome. *Nature genetics* **31**: 241–247.

Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**: 527–31.

Section 6: Subsampling analysis

Alexander Cagan, Christoph Theunert, Aida M. Andrés, Gabriel Santpere, Arcadi Navarro

Investigating the influence of sample size variation on results

The dataset analysed in this manuscript consists of population level whole-genome sequence data from multiple lineages. The number of individual genomes available varies between lineages, from three in the case of *G.b. graueri* to 12 for *P. paniscus* and *G.g gorilla*). This raises concerns that differences in sample size may influence our ability to detect signatures of selection in each lineage. This could lead to results that are due to differences in power rather than being biologically meaningful.

1. Methods

To assess the influence that variation in sample size has on our results we took a subsampling approach. For each of our selected tests we re-ran the entire analysis using only subsets of individuals. The overlap in the results obtained with these subsets could then be compared to the original results (using all individuals) to obtain a measure of how much variation in sample size is likely to be influencing our results.

For the HKA, FWH and MK tests we chose the *G.g. gorilla* and *P. paniscus* lineages for subsampling. We chose these lineages they have the largest sample sizes, providing the greatest opportunity to see the impact of sample size variation when subsampling. Coincidentally these lineages are also very different in their N_e (Figure 1), thus by subsampling in both lineages we may simulatenously explore whether differences in N_e may influence the effect of subsampling. For the ELS test we only use *G.g gorilla* for subsampling as *P. paniscus* is not included in the original ELS analysis.

1.1. HKA subsampling results

For the HKA we randomly selected four and eight individuals from *G.g. gorilla* and *P. paniscus*. This was repeated 100 times for each lineage and sample size respectively. For each of these subsamplings we re-ran the HKA analysis for chromosome one only and compared the results with the original results based on using all individuals. We analysed regions in the 0.001% and 0.01% tails for the positive and balancing selection tails respectively. For each region that occurred in a given tail on chromosome one in the original results we calculated the percentage of times it also appears in the tail in the subsamplings. So a region that remains in the tail in all 100 subsamplings has a percentage of 100%. The results are presented in supplementary figure S21.

The results suggest that for the HKA analyses variation in sample size does not have a major impact on the candidate positively selected regions. We observe that with the larger 0.01% tail there are a greater number of outlier regions that occur less frequently in the tail subsampling. However for both tails in *G.g Gorilla* and for the *P. paniscus* 0.01 tail for subsampling from either four or eight individuals the mean overlap of regions across subsamplings is almost 100%. The lowest mean overlap we observe is in the *P. paniscus* 0.01 tail when subsampling with eight individuals, where the mean overlap is 90%. It is unexpected that the overlap is lower when subsampling with eight individuals rather than four, however in both cases we consider the overall amount of overlap across subsamplings to be high. Therefore for the HKA positive selection tails we conclude that variation in sample size is not causing a strong bias in detecting putatively selected regions.

The results indicate that the balancing selection candidates for the HKA are much more influenced by sample size variation (supplementary figure S21). This may be due to the general rarity of balancing selection in the genome relative to positive selection. We expect only a small number of true targets of balancing selection on chromosome one, if any. Therefore the 'candidate' regions in the balancing selection tail on chromosome one may just be neutrally evolving regions, which may explain the low overlap between regions depending on sample size. Given that we detect the MHC region, a known target of balancing selection across a wide-range of vertebrates, in all our lineages despite their variation in sample size, we do not think that sample size variation is preventing us from detecting the strongest targets of balancing selection in the genome.

1.2 FWH subsampling results

We repeated the same subsampling analysis for FWH, though only for positive selection as this test does not detect signatures of balancing selection (supplementary figure S21). We observe that FWH is more sensitive to sample size variation than the HKA test for detecting candidate regions under positive selection. As expected, we observe a greater overlap with the original results when using the subsampling of eight individuals compared to when we use only four (supplementary figure S22(E-J)). We also observe higher overlap with the original results with the more stringent 0.001% tail compared to the 0.01% tail, suggesting that the strongest candidates of positive selection are relatively robust to subsampling. We also observe subsampling results in a higher mean percentage of overlaps with the original results with *P. paniscus* compared to *G.g. gorilla* [11], although the ranges overlap.

The differing sensitivity to sample size variation between these two tests may be due to the different types of information that they use to detect signatures of selection. The HKA test is dependent on the ratio of substitutions to polymorphisms, without considering the allele frequency of polymorphisms. This should make it particularly robust to sample size variation. In contrast, methods that are more reliant on fluctuations in the allele frequency spectrum to detect selection, such as FWH, are more sensitive to variation in sample

size. Therefore, our results provide support for using the HKA test to detect selection in cases where sample size is low or unequal between populations.

1.3 ELS subsampling results

We performed a similar subsampling analysis for the ELS test, randomly subsampling four or eight randomly selected individuals from *G.g. gorilla* 100 times in both cases. We find that overlap of candidate selected regions between the original results and the subsamplings is generally low, with a mean overlap of ~20% for each putatively selected region across subsamplings. This may be due to the reliance of ELS on using simulations to infer parameters, which makes the test vulnerable to errors if the demographic model in the simulations is inappropriate.

1.4 HKA and Ne correlations subsampling results

We observe significant correlations between the long term Ne of lineages and the percentage of protein coding exons in several bins of the HKA empirical distribution (Supplementary Materials 2.2). To ensure that these correlations are not due to variance in sample size between lineages we re-ran these correlations after excluding lineages with either high or low sample-sizes. We re-ran the correlations using only populations with < 10 individuals (excluding *G.g gorilla*, *P. paniscus* and *P.t. ellioti*) and only populations with > 5 individuals (excluding *P. abelii*, *P. pygmaeus*, *P.t. troglodytes* and *Pt. Verus*). In both cases there were at least five lineages.

The results for both sub-sampled correlations are very similar to the original results with all lineages (supplementary table S106, supplementary fig S22). The subsampling excluding lineages with larger sample sizes has the most similar correlation scores to our original results (supplementary table S106). However the R values are very similar to the original results in both cases (supplementary figure S22). This suggests that the correlation between the long-term Ne of a lineage and the percentage of protein-coding exons in bins of the HKA empirical distribution, which we infer as a measure of the strength of background selection, is robust to variation in sample size between lineages.

Section 7: Targets of selection

Alexander Cagan, Christoph Theunert, Jaume Bertranpetit, Aida M. Andrés

1. Introduction

Here we provide further discussion of genes and pathways with signatures of selection identified by our analyses which were not presented in the main Discussion section.

1.1. Diet

All lineages of the *Hominidae* are omnivorous, with all lineages apart from *H. sapiens* having a preference for frugivory (Boyd & Silk, 1997). However, there is still considerable variation in diet between lineages (Uchida, 1996). Adaptations that maximize the extraction of energy from food are expected to be highly beneficial. As a result we might expect genes related to digestive processes to have been targeted by positive selection.

Human populations tend to consume high levels of starch rich foods compared to other members of the *Hominidae* (Hohmann et al. 2012), a trend exacerbated by the transition to agriculture and the widespread availability of starchy foods (Zohary et al. 2012). Previous work identified copy number variation in *AMY1* in humans as advantageous to increase the benefits from starchy foods (Perry et al. 2007). We provide additional evidence that genes related to starch metabolism have been targets of positive selection in humans. In *H. sapiens* the strongest category enrichment among the HKA candidate targets of positive selection is for the KEGG pathway 'starch and sucrose metabolism' (p-value=0.02), a signal not shared among any of the other ape lineages. This signal is driven by the genes *GAA*, *AMY2B* and *GUSB*. Incidentally, copy number increase of *AMY2B* in dogs is considered an adaptation to starch-rich diets that arose during cohabitation with humans (Axelsson et al. 2013). The evidence of positive selection on this gene in humans suggests that adaptation to a starch-rich diet may have a partially shared genetic basis between dogs and humans.

1.2 Anatomy

There is considerable anatomical variation between the different lineages of the *Hominidae*. One of the most clearly visible differences between lineages is in body size, with gorillas being the largest extant primate. The ELS test detects signatures of positive selection in *G.g. gorilla* in a region containing the gene *IGF2R*, which encodes insulin growth factor receptor 2. Mutations in this gene are associated with variation in body size traits in cattle (Berkowicz et al. 2012). Biological category enrichment analysis also shows significant enrichment on the KEGG pathway 'vascular smooth muscle contraction' (p-value=0.03). Changes in the vascular system, which regulates blood pressure, may have been necessary to cope with the changes in body size that must have occurred during evolution of the gorilla lineage. Therefore it is possible that

selection on these genes are related to body size differences between the *Gorilla* subspecies, with *G.g. gorilla* considered to have less sexual dimorphism in body size and a smaller male body size than *G.g. beringei* (Taylor 1997).

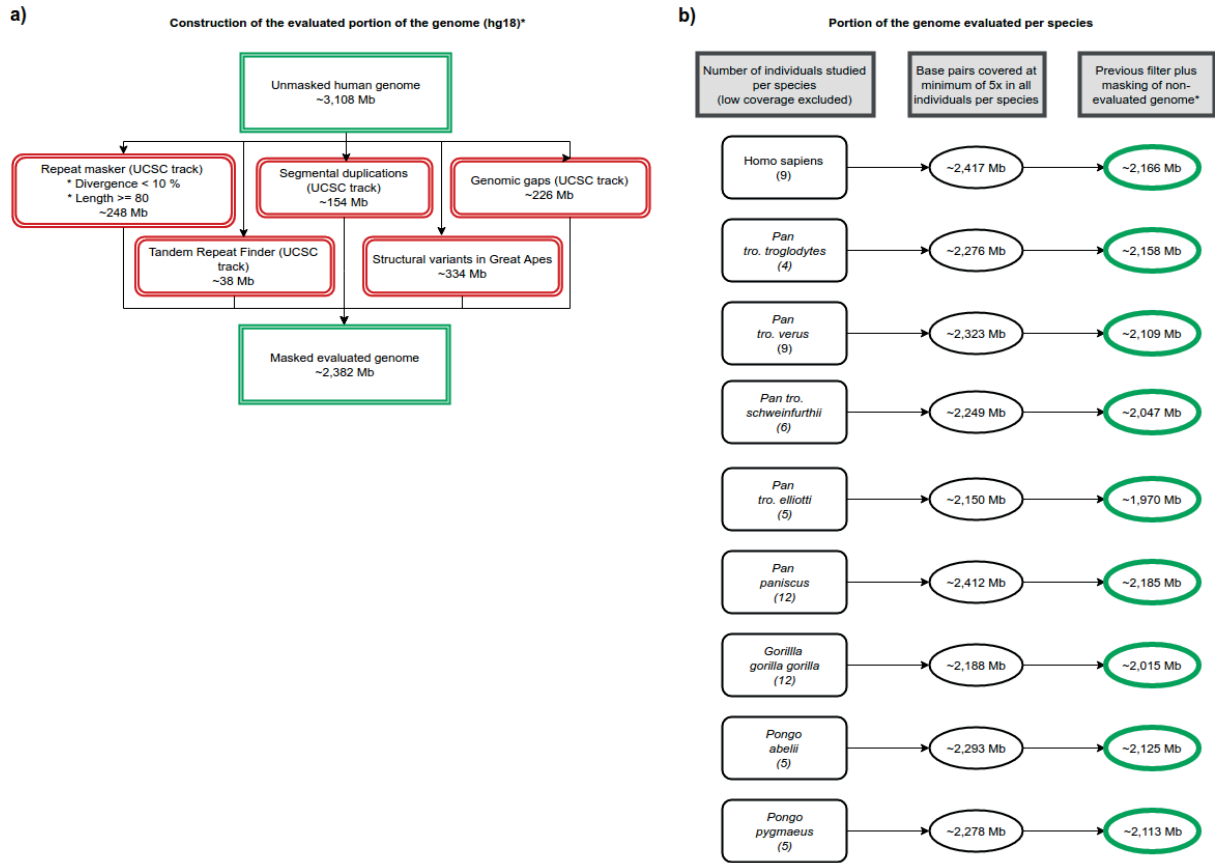
The *Hominidae* lineages display a variety of anatomical adaptations to their various forms of quadrupedal and bipedal locomotion. *Pongo* are the only lineage in the *Hominidae* that regularly brachiate (a form of arboreal locomotion based on swinging through trees using only the arms), which likely involved some molecular adaptations. Interestingly, in *P. abelii* HKA candidate targets of positive selection are significantly enriched for genes in the GO molecular function category 'structural constituent of muscle' (p-value=0.01). One of the genes driving this signal is *NEB*, which encodes nebulin, a protein that helps to maintain the structural integrity of myofibrils in skeletal muscle. Deficits in nebulin result in a dramatic decrease in the force production capacity of skeletal muscle (Bang et al. 2006). This gene also shows a signature of positive selection from the HKA test in *P. abelii* and also *P.t. verus*, as well as a recent signature of positive selection in *P. abelii* from the FWH test.

Another feature differentiating the *Hominidae* lineages is the morphology and distribution of their body hair (Yesudian 2011). However, the genetic basis of such differences is not well understood. Genes related to 'hair follicle' development were identified as accelerated in the gorilla lineage in a recent study that used dN/dS ratios to compare protein coding sequences in humans, chimpanzees and gorillas (Scally et al. 2012). One of the genes they identified as contributing to this signal is *DSG4*. We find that this gene appears particularly constrained in all species of the genus *Pan* (MK results) but it shows signatures of positive selection in *G.g. gorilla*. This gene encodes the protein desmoglein 4, which plays an important role in the maintenance of hair follicle keratinocytes (Bazzi et al. 2006). Mutations in *DSG4* can cause hypotrichosis in humans, which is an abnormal condition of hair that affects its amount and results in atrophied hair follicles and shafts (Shimomura et al. 2006).

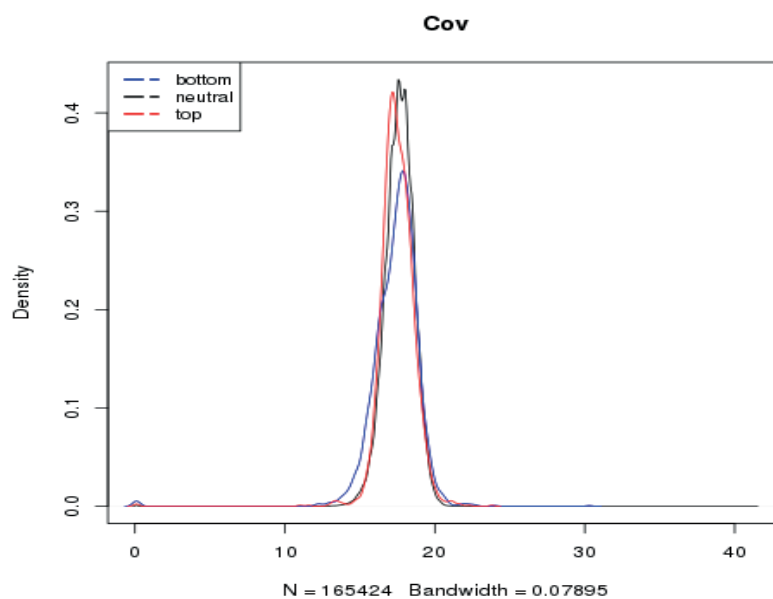
References

- Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**: 360–4.
- Bazzi H, Getz A, Mahoney MG, Ishida-Yamamoto A, Langbein L, Wahl JK, Christiano AM. 2006. Desmoglein 4 is expressed in highly differentiated keratinocytes and trichocytes in human epidermis and hair follicle. *Differentiation* **74**: 129–140.
- Berkowicz EW, Magee DA, Berry DP, Sikora KM, Howard DJ, Mullen MP, Evans RD, Spillane C, MacHugh DE. 2012. Single nucleotide polymorphisms in the imprinted bovine insulin-like growth fac-

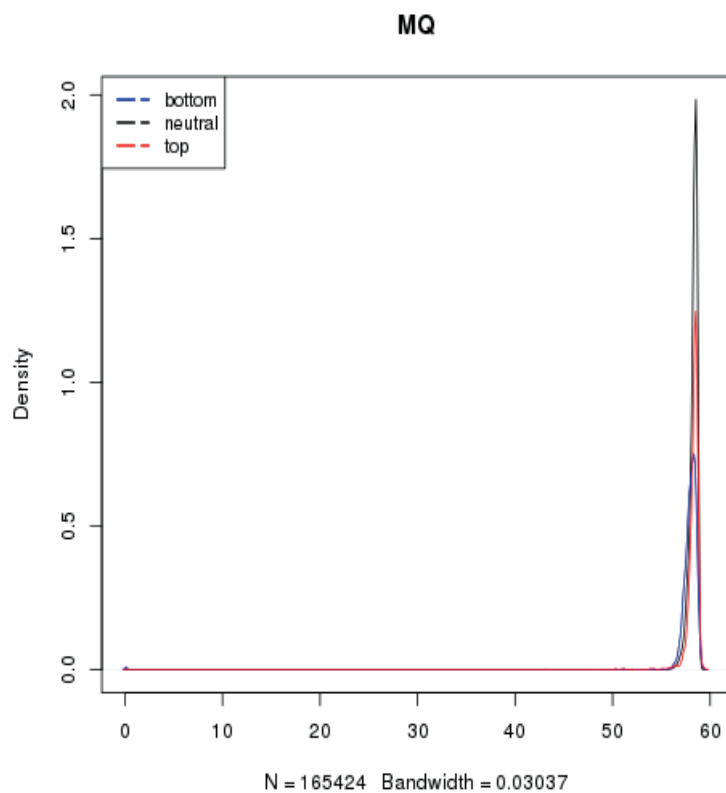
- tor 2 receptor gene (IGF2R) are associated with body size traits in Irish Holstein-Friesian cattle. *Animal genetics* **43**: 81–7.
- Boyd R, Silk J. 1997. *How Humans Evolved*. W.W. Norton & Company, New York.
- Hohmann G, Robbins MM, Boesch C. 2012. *Feeding Ecology in Apes and Other Primates*. Cambridge University Press.
- Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nature genetics* **39**: 1256–1260.
- Scally A, Dutheil J, Hillier L. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169–175.
- Shimomura Y, Sakamoto F, Kariya N, Matsunaga K, Ito M. 2006. Mutations in the desmoglein 4 gene are associated with monilethrix-like congenital hypotrichosis. *The Journal of investigative dermatology* **126**: 1281–5.
- Taylor AB. 1997. Relative growth, ontogeny, and sexual dimorphism in gorilla (*Gorilla gorilla gorilla* and *G. g. beringei*): evolutionary and ecological considerations. *American journal of primatology* **43**: 1–31.
- Uchida A. 1996. What we don't know about great ape variation. *Trends in ecology and evolution* **11**: 163–168.
- Zohary D, Hopf M, Weiss E. 2012. *Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin*. OUP Oxford.



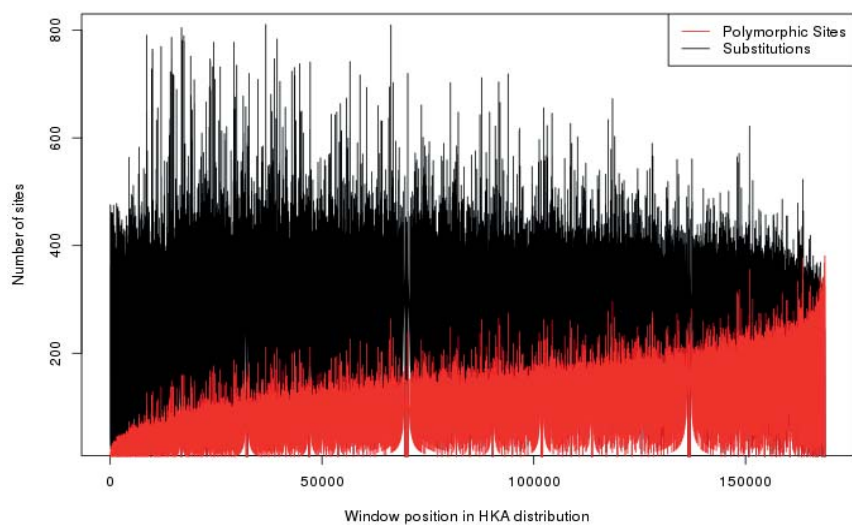
Supplementary Figure S1. Flowchart of filtering steps. **A.** Filtering steps that were applied across the entire dataset. Each red box represents a particular filtering step and the amount of the genome (Mb) that was excluded. Green boxes show the amount of the genome available for analysis before and after these filtering steps. **B.** Portion of the genome evaluated in each species, based on filtering of sites with < 5x coverage across all individuals per species



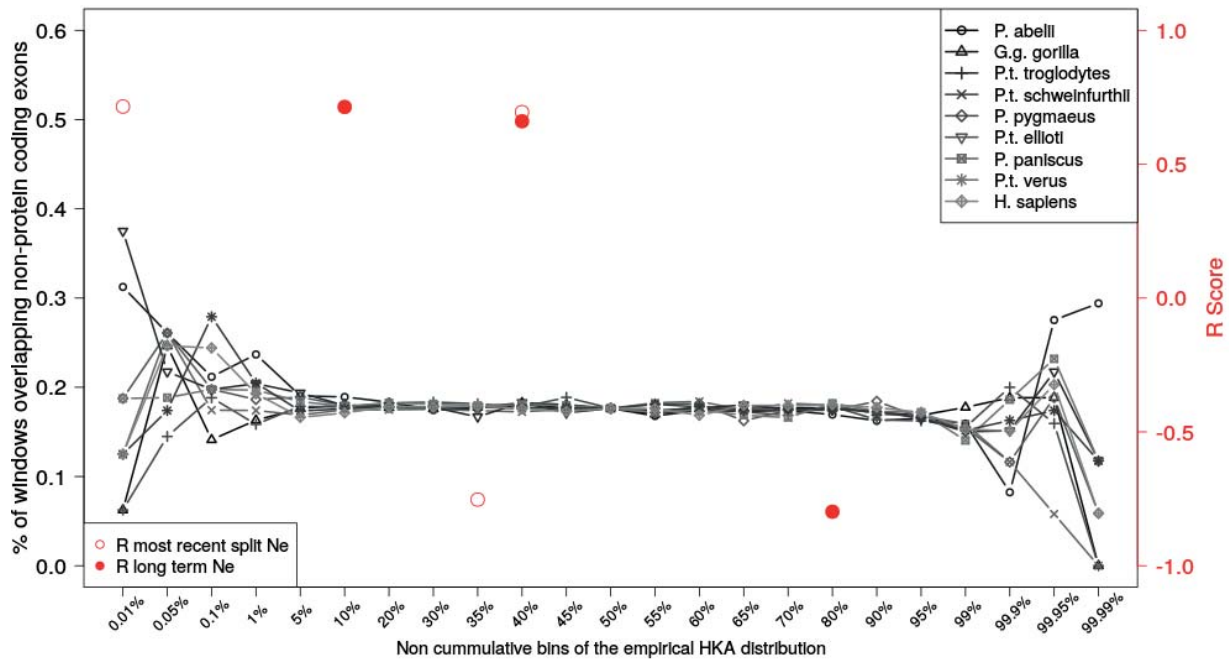
Supplementary Figure S2. Plot showing relationship between coverage and HKA score for *G.g. gorilla*. The X-axis shows average coverage. The Y-axis shows the frequency. Red lines represent results from the top 1% of the HKA score distribution, blue lines from the bottom 1% and black lines from the middle 98%. The vertical bars represent the mean coverage after 1000 permutations from each distribution.



Supplementary Figure S3. Plot showing relationship between the Mapping Quality Score and HKA score for *G.g. gorilla*. The X-axis represents the mean Mapping Quality Score in each 30kb window. The Y-axis represents the frequency. Red lines represent results from the top 1% of the HKA score distribution, blue lines from the bottom 1% and black lines from the middle 98%. The vertical bars represent the mean coverage after 1000 permutations from each distribution.

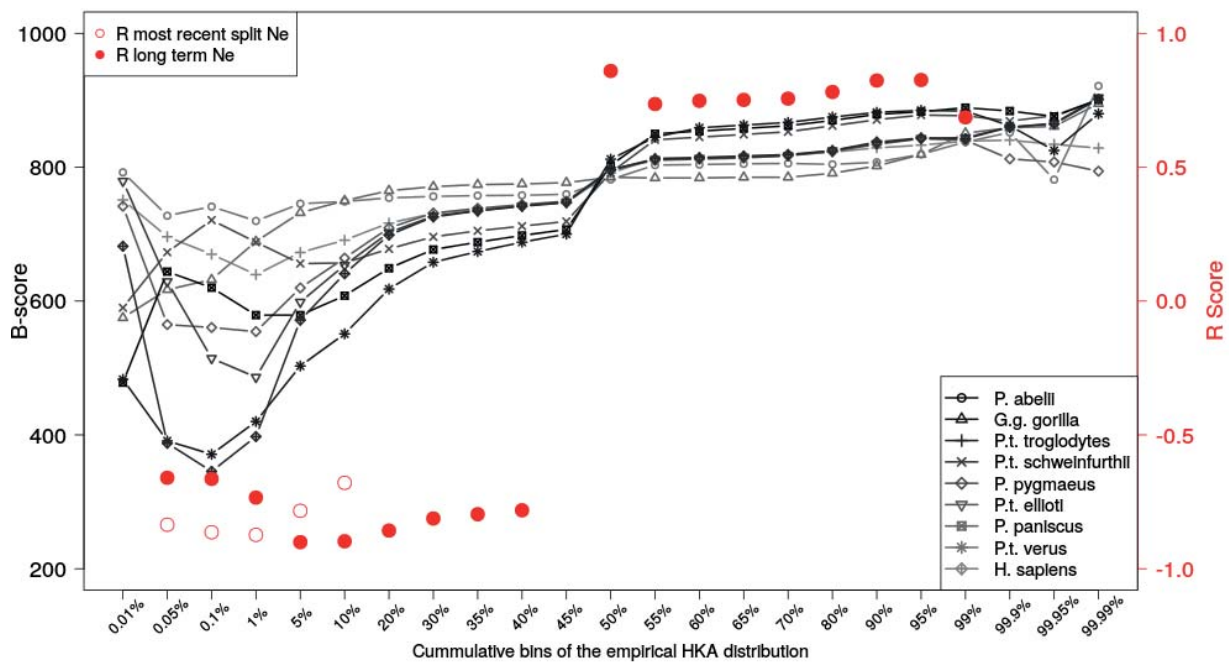


Supplementary Figure S4. Number of substitutions and polymorphic sites per window across the HKA empirical distribution for *P. troglodytes*. The Y-axis shows the number of sites in a window. The X-axis shows the position of windows in the HKA empirical distribution. The HKA score of windows increases along the X-axis.



Supplementary Figure S5. Percentage of windows overlapping non-protein coding exons.

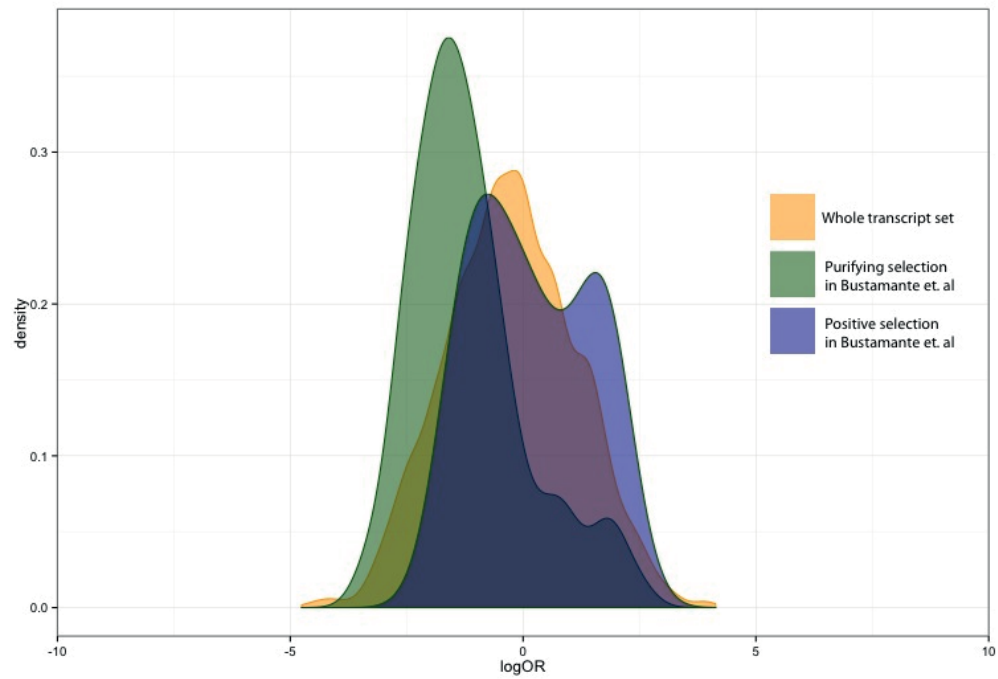
Percentage of windows overlapping non-protein coding exons (≥ 1 bp overlap exon with genomic window) for non-cumulative bins of the HKA empirical distribution (X-Axis). We plot this for each lineage as a shaded line. Furthermore, for each bin we measure the correlation between the % of windows overlapping non-protein coding exons and N_e among all lineages using a Pearson's correlation analysis. We do this separately with an estimate of short-term and long-term N_e , derived from PSMC and Watterson's estimator respectively (taken from Prado-Martinez *et al.* 2013). The right-side Y-axis shows the R score. R score is plotted with dashed lines. Only R values with significant P values ($P < 0.05$) are shown.



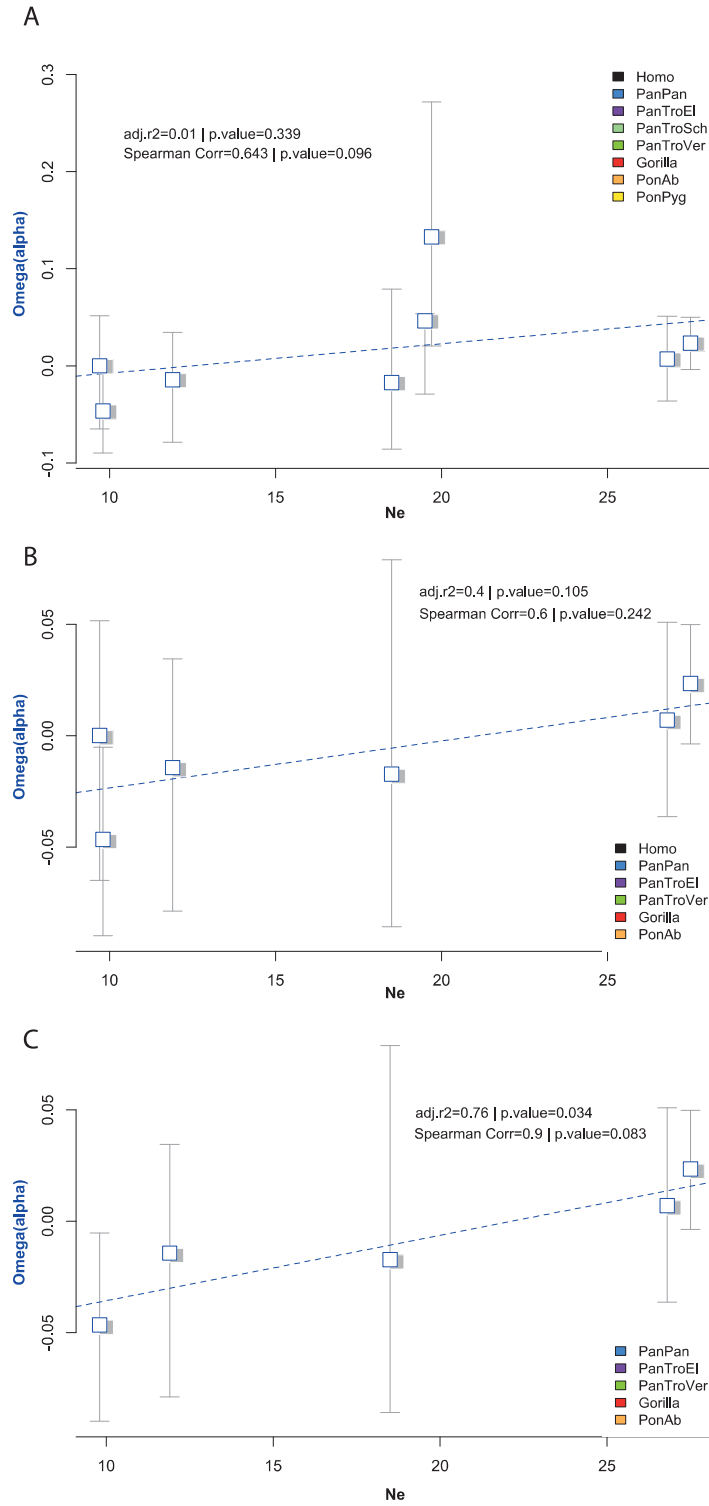
Supplementary Figure S6. Average B-score for windows with different cut-offs of the HKA empirical distribution. Average B-scores for each cumulative bin of the HKA empirical distribution (X-axis). B-scores were calculated according to McVicker *et al.* (2009). We plot this for each lineage as a shaded line. Furthermore, for each bin we measure the correlation between the average B-score and N_e among all lineages using a Pearson's correlation analysis. We do this separately with an estimate of short-term and long-term N_e , derived from PSMC and Watterson's estimator respectively (taken from Prado-Martinez *et al.* 2013). The right-side Y-axis shows the R score. R score is plotted with dashed lines. Stars indicate R scores with significant P values ($P < 0.05$).

See file 'Supplementary_Figures_S7_S15.pdf' for Supplementary Figures S7-S15

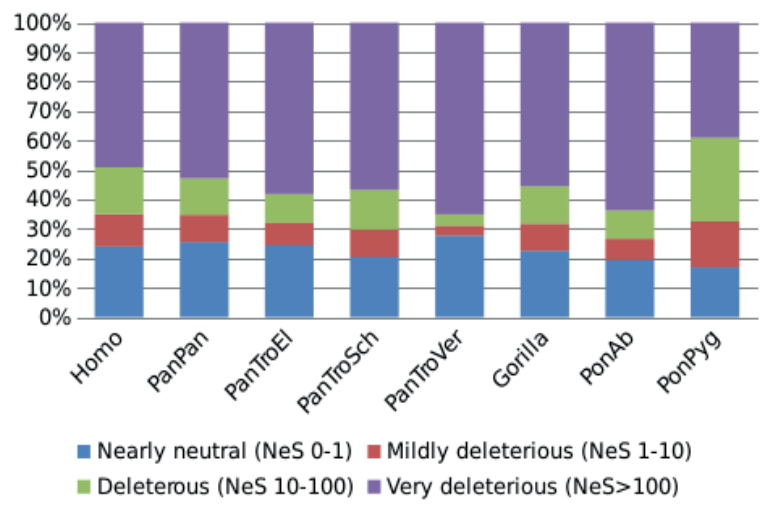
Supplementary Figures S7-S15. Annotation in HKA 0.1% tail compared to neutral sub-sampling. For each lineage we separately calculated the percentage of windows in the 0.1% tail of the HKA distribution containing any type of functional annotation, protein coding exons, or non-protein coding exons (see title of figure for category presented). These results were compared to 100 random sub-samplings of an equal number of windows from the genome-wide distribution. The value from the 0.1% tail is given as a horizontal line while the results from the neutral sub-samplings are given as a box-plot. P-values indicating whether the percentage of annotation in the 0.1% tail is significantly enriched relative to the genome-wide sub-sampling are presented below the X-axis.



Supplementary Figure S16. DFE of deleterious mutations in 0-fold sites. Density plot of MK-test logOR for genes showing different signals of selection in Bustamante et al., 2005.

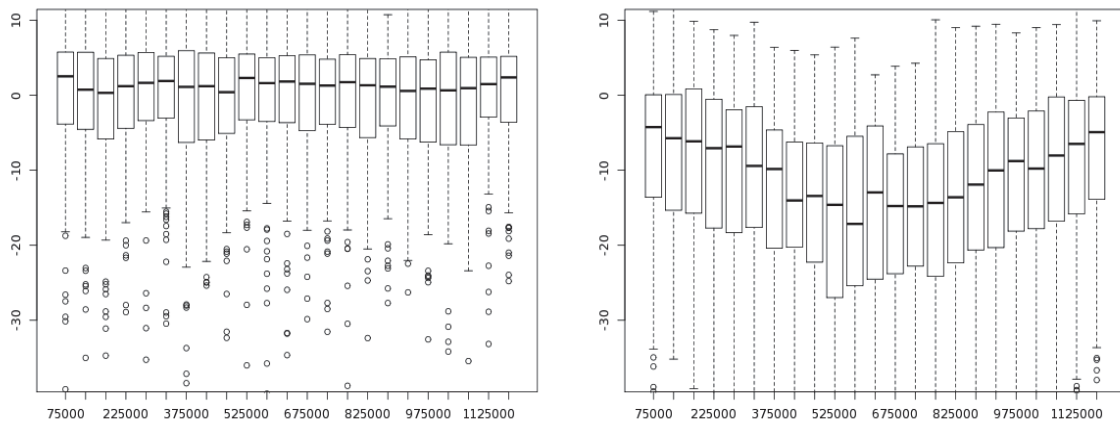


Supplementary Figure S17. Correlation between rate of adaptive substitutions (ω_a) and effective population size (N_e), using all species (A), excluding *P.t.schweinfurthii* and *P. abelii* (B), also excluding *H. sapiens* (C).



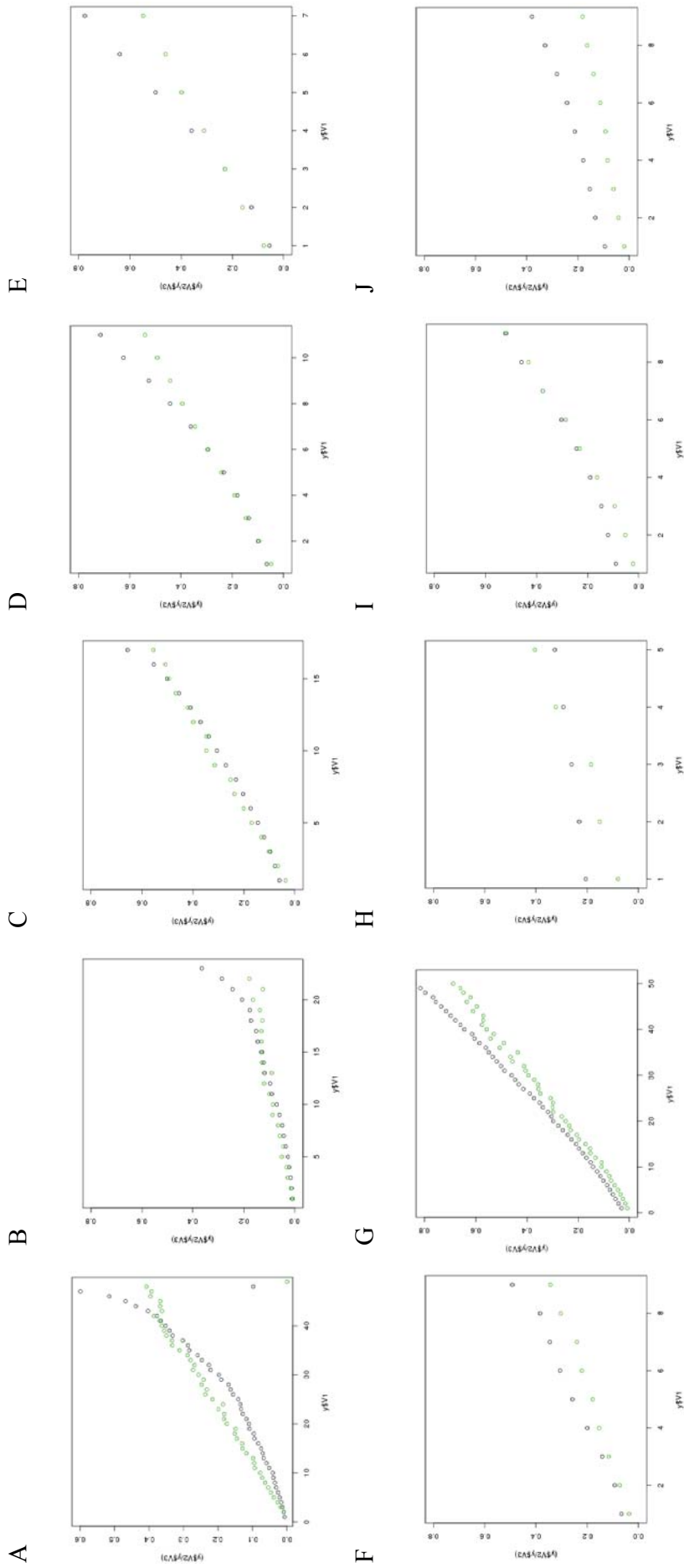
Supplementary Figure S18. Distribution of fitness effect of mutations at 0-fold sites.

FayWu_EUR_MAF0.0001_winLength30000_offset3000_minSNPs1_ptailN.faywu.FayWu



Supplementary Figure S19. Validation of Fay and Wu's H algorithm with simulations.

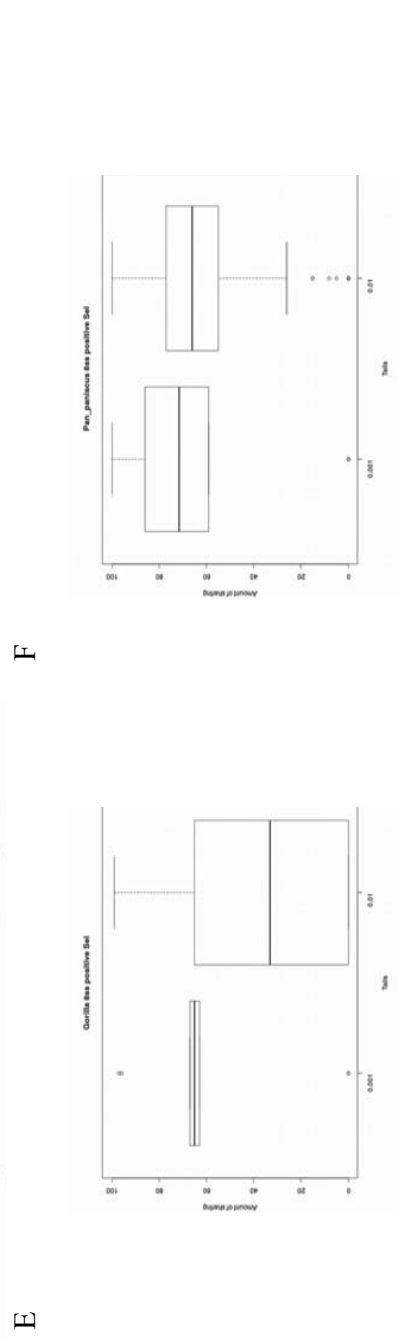
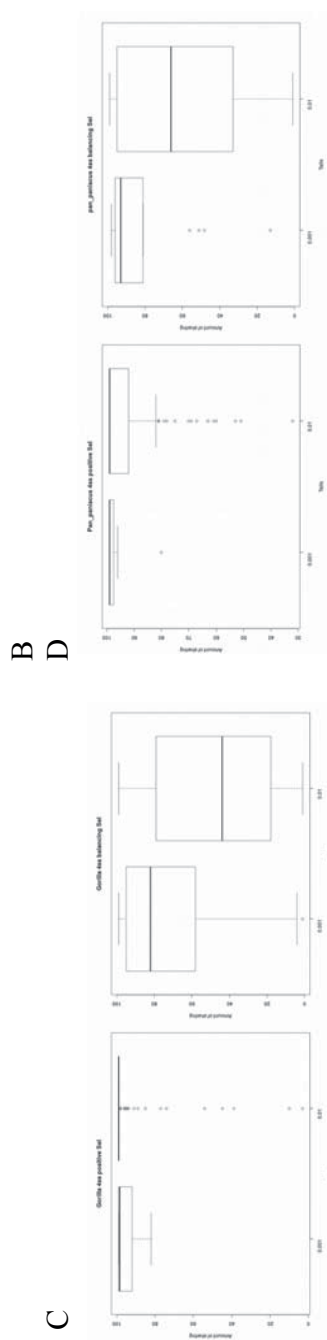
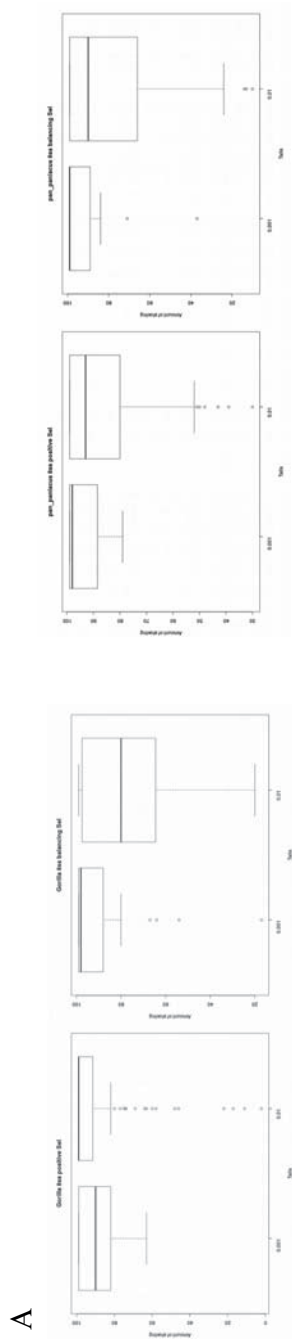
Validation of FWH algorithm in a neutral simulation (left) and with a selective event in the central region (right). COSI simulator (Schaffner et 2005) was run under the best fit model (validated human demography) to generate 1000 realistic simulations under neutrality and under selection. Simulations with selection consisted in a selective sweep starting 500 generations ago with a selection coefficient of 0.022 located in the middle position of the 1.2Mb simulation. Both scenarios were then analyzed using Fay Wu's H algorithm. Plots were generated averaging the obtained scores in windows of 25Kb.



Supplementary Figure S20. Pairwise comparison between 2 analyzed lineages, *species 1* – *species 2*. Each graph shows the fraction of sites where *species 1* is derived in dependence of number of individuals in *species 2*. Simulated results in green, observed results in black.

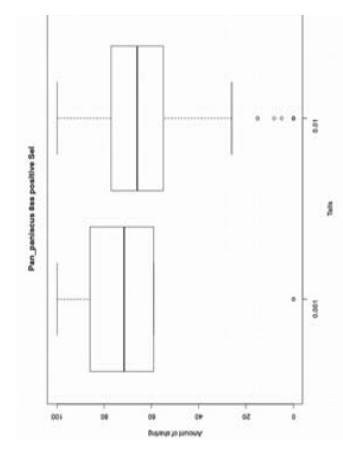
A. *P. troglodytes* – *P. paniscus*. **B.** *P. paniscus* – *P. troglodytes*. **C.** *P.t. ellioti* – *P. paniscus*. **D.** *P.t. schweinfurthii* – *P. paniscus*. **E.** *P.t. troglodytes* – *P. paniscus*. **F.** *P.t.*

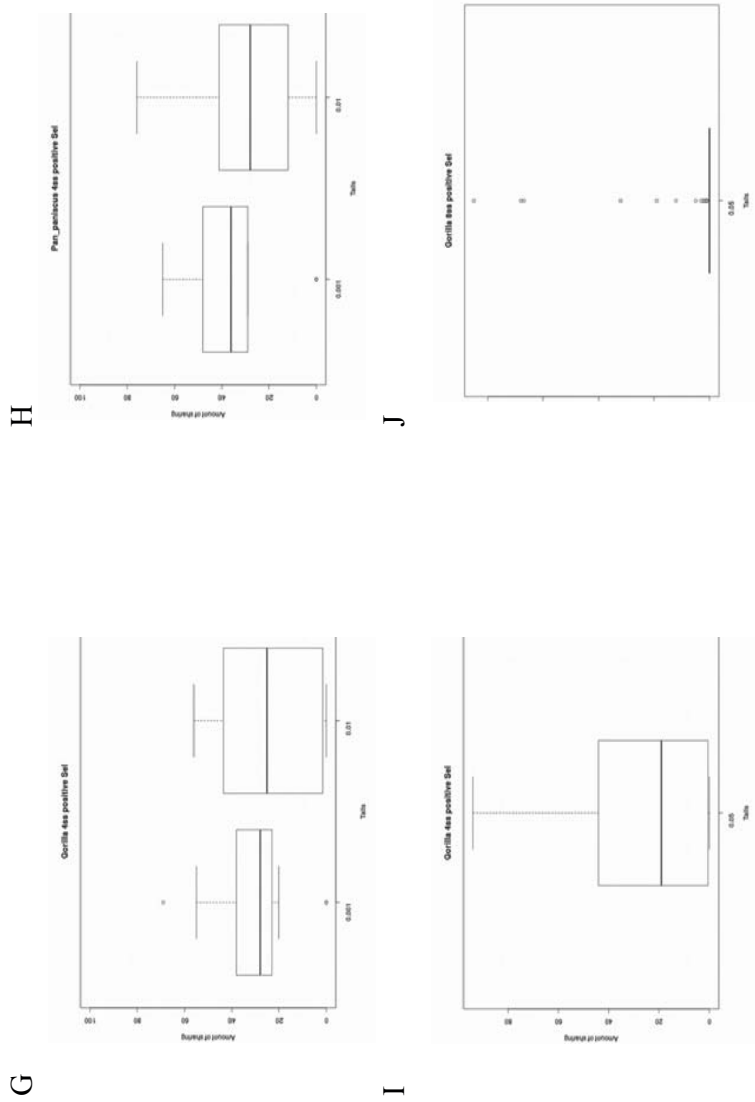
versus *P. paniscus*. **G.** *G.g. gorilla* – *G.b. graueri*. **H.** *G.b. graueri* – *G.g. gorilla*. **I.** *P. abelii* – *P. pygmaeus*. **J.** *P. pygmaeus* – *P. abelii*.



D

E



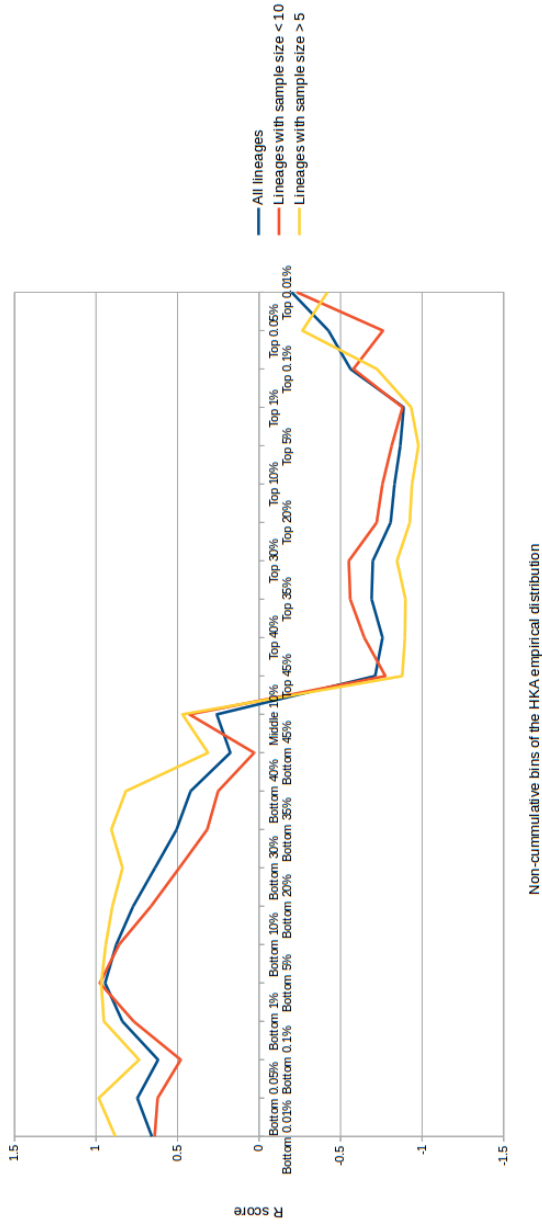


Supplementary Figure S21. Subsampling results for HKA, FWH and ELS with 4 and 8 individuals for *G.g. gorilla* and *P. paniscus*.

Box plots summarizing the amount of sharing of candidate regions from the original test results with the 100 subsampling analyses. For HKA test, the left-side plot presents results for the positive selection tail. The right-side plot presents results for the balancing selection tail. The X-axis shows the fraction of regions from the empirical distribution that are considered in the tail (far-left tail for positive selection, far-right for balancing selection). The Y-axis shows the percentage of overlap between regions from the original results compared to the 100 subsamplings. For FWH, Box plots summarizing the amount of sharing of candidate regions from the original test results on positive selection with the 100 subsampling analyses. For ELS, Box plots summarizing the amount of sharing of the top 5% of putatively

selected regions from the original test results with the 100 subsampling analyses.

- A. Subsampling results for HKA with 8 individuals for *G.g. gorilla*.
- B. Subsampling results for HKA with 8 individuals for *P. paniscus*.
- C. Subsampling results for HKA with 4 individuals for *G.g. gorilla*.
- D. Subsampling results for HKA with 4 individuals for *P. paniscus*.
- E. Subsampling results for FWH with 8 individuals for *G.g. gorilla*.
- F. Subsampling results for FWH with 8 individuals for *P. paniscus*.
- G. Subsampling results for FWH with 4 individuals for *G.g. gorilla*.
- H. Subsampling results for FWH with 4 individuals for *P. paniscus*.
- I. Subsampling results for ELS with 4 individuals.
- J. Subsampling results for ELS with 8 individuals.



Supplementary Figure S22. Subsampling results for HKA Ne correlations. Correlations between Watterson's Ne estimates for each lineage and the percentage of regions overlapping protein-coding exons in non-cumulative bins of the HKA empirical distribution. Results from using all lineages are plotted alongside results calculated using only lineages with a sample size < 10 or > 5 . The X-axis shows the non-cumulative bins of the HKA empirical distribution. The Y-axis shows the Pearson correlation (R) between the E statistic and Ne within each HKA bin and across all lineages.

References

- [1] Darwin C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle For Life*. London (United Kingdom). John Murray.
- [2] Kimura M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press, 1983.
- [3] Nielsen R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.*, 39, 197-218.
- [4] Smith JM, & Haigh J. (1974). The hitch-hiking effect of a favourable gene. *Genetical research*, 23(01), 23-35.
- [5] Pritchard JK & Di Rienzo A. (2010). Adaptation—not by sweeps alone. *Nature Reviews Genetics*, 11(10), 665-667.
- [6] Nielsen R, Hellmann I, Hubisz M, Bustamante C, & Clark AG. (2007). Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, 8(11), 857-868.
- [7] Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, & Laval G. (2014). Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Molecular biology and evolution*, 31(7), 1850-1868.
- [8] Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, ... & Simianer H. (2014). Classic selective sweeps revealed by massive sequencing in cattle. *PLoS Genetics*, 10(2), e1004148.
- [9] 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68-74.
- [10] Diamond J. (1999) *Guns, Germs, and Steel: The Fates of Human Societies*. Norton & Company.
- [11] Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, ... & Doust AN. (2014). Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences*, 111(17), 6139-6146.
- [12] Clutton-Brock J. (1992). The process of domestication. *Mammal Review*, 22(2), 79-85.
- [13] Andersson L, & Georges M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics*, 5(3), 202-212.

- [14] Frantz LA, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, ... & Tresset A. (2016). Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, 352(6290), 1228-1231.
- [15] Clutton-Brock J. (1995). Origins of the dog: domestication and early history. *The domestic dog: Its evolution, behaviour and interactions with people*, 7-20.
- [16] Evans LT. (1984). Darwin's use of the analogy between artificial and natural selection. *Journal of the History of Biology*, 17(1), 113-140.
- [17] Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, ... & Novembre J. (2014). Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genetics*. 10, e1004016.
- [18] Davis SJM, Valla FR. (1978) Evidence for domestication of the dog 12,000 years ago in the Natufian of Israel. *Nature*. 276:608–10
- [19] Wilkins AS, Wrangham RW, & Fitch WT. (2014). The “domestication syndrome” in mammals: A unified explanation based on neural crest cell behavior and genetics. *Genetics*, 197(3), 795-808.
- [20] Sánchez-Villagra MR, Geiger M, & Schneider RA. (2016). The taming of the neural crest: a developmental perspective on the origins of morphological covariation in domesticated mammals. *Royal Society open science*, 3(6), 160107.
- [21] Belyaev D. (1979) Destabilizing selection as a factor in domestication. *J. Hered.* 70:301-8
- [22] Trut L, Oskina I, Kharlamova A. (2009) Animal evolution during domestication: the domesticated fox as a model. *BioEssays*. 31: 349–360.
- [23] Albert FW, Shchepina O, Winter C, Römppler H, Teupser D, Palme R, ... & Pääbo S. (2008). Phenotypic differences in behavior, physiology and neurochemistry between rats selected for tameness and for defensive aggression towards humans. *Hormones and Behavior*, 53(3), 413-421.
- [24] Albert FW, Carlborg Ö, Plyusnina I, Besnier F, Hedwig D, Lautenschläger S, ... & Pääbo S. (2009). Genetic architecture of tameness in a rat model of animal domestication. *Genetics*, 182(2), 541-554.
- [25] Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, ... & Sninsky JJ. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, 3(6), e170.

- [26] Hobolth A, Dutheil JY, Hawks J, Schierup MH, & Mailund T. (2011). Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome research*, 21(3), 349-356.
- [27] Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, ... & McCarthy S. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388), 169-175.
- [28] Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, ... & Pääbo S. (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature*, 486(7404), 527-531.

Acknowledgements

First and foremost I would like to express my gratitude to Svante Pääbo for his supervision and guidance. Without his support none of this would have been possible. Most importantly, he taught me to think like a scientist. This project has taken me around the world and expanded virtually all of my horizons. I could not have asked for a better mentor.

To Aida M. Andrés, for instilling in me her enthusiasm for population genetics. Her positivity and support made work here a pleasure.

To Frank W. Albert, for his support throughout my time here, particularly in the early days when I was finding my feet.

To the Wednesday (and later Thursday) group, for providing a stimulating environment to share and discuss ideas. During these meetings I learned how to collect, present and evaluate data. I hope to keep the spirit of the group with me wherever I go.

To my collaborators in Akademgorodok, who always welcomed me and made Novosibirsk feel like home, no matter the weather. Allowing me to share in this adventure with you is something I will always cherish. In particular, my thanks to Irina Plyusnina, whose passion for life and for research I carry with me. You are greatly missed.

To Heike and Sandra, who welcomed me when I could barely speak a word of German and had never even seen a rat cage before. Their friendship has made life and work here a much richer experience.

To all the collaborators who contributed their time and experience to these projects and made them possible.

To all my colleagues at the MPI. There are too many wonderful people to mention but you know who you are. I can't thank you enough for your friendship, support, and chocolate. I will miss you all.

To all the staff at the MPI who have supported me. To Rigo and Ines for their assistance on countless occasions. To Rocco for helping me get settled. I would like to especially thank Viola, for her miraculous ability to make everything run so smoothly and to create order out of my chaos.

A very special thank you to my family and to Harriet for their endless support and tolerance.

Finally, to the rats and mink that gave their lives so that we might see further.

Curriculum Vitae

Alex Cagan

31 Hardenberg Strasse
Leipzig 04275
Germany
+49 15170812594
alexander_cagan@eva.mpg.de

Academic Vita

- 2010-Present PhD Evolutionary Genetics
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
Supervisor: Svante Pääbo
Thesis: *Comparative genomic approaches to human evolutionary history*
Also participate in research to detect positive selection in the Great Apes
- 2013 Master of Arts
St Catharine's College, University of Cambridge, Cambridge, UK
- 2006-2009 B.A. Archaeology & Anthropology
St Catharine's College, University of Cambridge, Cambridge, UK
Part II Biological Anthropology: 1st class with distinction (highest honours)
Part I Archaeology & Anthropology: 1st class

Publications

Alex Cagan, Christoph Theunert, Hafid Laayouni, Gabriel Santpere, Marc Pybus, Ferran Casals, Kay Prüfer, Arcadi Navarro, Tomas Marques-Bonet, Jaume Bertranpetit, Aida M. Andrés. "Natural selection in the great apes" *Molecular Biology and Evolution* (2016) doi: 10.1093/molbev/msw215

Alex Cagan & Torsten Blass, " Identification of genomic variants putatively targeted by selection during dog domestication". *BMC Evolutionary Biology* 16:10 (2016) doi: 10.1186/s12862-015-0579-7

Alex Cagan, Frank W Albert, Irina Plyusnina, Lyudmila Trut, Gabriel Renaud, Frederic Romagné, Victor Wiebe, Rimma Kozhemjakina, Rimma Gulevich, Oleg Trapezov, Nikolay Yudin, Tatyana Alekhina, Ruslan Aitnazarov, Ludmila Trapezova, Yury Herbeck, Torsten Schöneberg, Svante Pääbo. "Genes and pathways selected during animal domestication". *Manuscript submitted to eLife*.

Henrike Heyne, Susann Lautenschläger, Ronald Nelson, François Besnier, Maxime Rotival, **Alex Cagan**, Rimma Kozhemyakina, Irina Z. Plyusnina, Lyudmila Trut, Örjan Carlborg, Enrico Petretto, Leonid Kruglyak, Svante Pääbo, Torsten Schöneberg, Frank W. Albert. "Genetic influences on brain gene expression in rats selected for tameness and aggression". *Genetics* 198, no. 3 (2014): 1277-1290.

Frantz, Laurent AF, Joshua Schraiber, Ole Madsen, Hendrik-Jan Megens, **Alex Cagan**, Mirte Bosse, Yogesh Paudel, Richard PMA Crooijmans, Greger Larson, and Martien AM Groenen. "Analyses of Eurasian wild and domestic pig genomes reveals long-term gene-flow during domestication." *Nature genetics* 47 (2015): 1141-1148.

Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, August E. Woerner, Timothy D. O'Connor, Gabriel Santpere, **Alex Cagan**, Christoph Theunert, Ferran Casals, Hafid Laayouni, Kasper Munch, Asger Hobolth, Anders E. Halager, Maika Malig, Jessica Hernandez-Rodriguez, Irene Hernando-Herraez, Kay Prüfer, Marc Pybus, Laurel Johnstone, Michael Lachmann, Can Alkan, Dorina Twigg, Natalia Petit, Carl Baker, Fereydoun Hormozdiari, Marcos Fernandez-Callejo, Marc Dabad, Michael L. Wilson, Laurie Stevison, Cristina Camprubí, Tiago Carvalho, Aurora Ruiz-Herrera, Laura Vives, Marta Mele, Teresa Abello, Ivanela Kondova, Ronald E. Bontrop, Anne Pusey, Felix Lankester, John A. Kiyang, Richard A. Bergl, Elizabeth Lonsdorf, Simon Myers, Mario Ventura, Pascal Gagneux, David Comas, Hans Siegismund, Julie Blanc, Lidia Agueda-Calpena, Marta Gut, Lucinda Fulton, Sarah A. Tishkoff, James C. Mullikin, Richard K. Wilson, Ivo G. Gut, Mary Katherine Gonder, Oliver A. Ryder, Beatrice H. Hahn, Arcadi Navarro, Joshua M. Akey, Jaume Bertranpetit, David Reich, Thomas Mailund, Mikkel H. Schierup, Christina Hvilsom, Aida M. Andrés, Jeffrey D. Wall, Carlos D. Bustamante, Michael F. Hammer, Evan E. Eichler & Tomas Marques-Bonet. "Great ape genetic diversity and population history." *Nature* 499, no. 7459 (2013): 471-475.

Andrea Bamberg Migliano, Irene Gallego Romero, Mait Metspalu, Matthew Leavesley, Luca Pagani, Tiago Antao, Da-Wei Huang, Brad T. Sherman, Katharine Siddle, Clarissa Scholes, Georgi Hudjashov, Elton Kaitokai, Avis Babalu, Maggie Belatti, **Alex Cagan**, Bryony Hopkinshaw, Colin Shaw, Mari Nelis, Ene Metspalu, Reedik Mägi, Richard A. Lempicki, Richard Villems, Marta Mirazon Lahr, & Toomas Kivisild. "Evolution of the Pygmy Phenotype: Evidence of Positive Selection from Genome-wide Scans in African, Asian, and Melanesian Pygmies." *Human Biology* 85:1-3 (2013): 251-284

Awards

Walter M. Fitch Award for best presentation at the Walter M. Fitch symposium at the annual meeting of the Society for Molecular Biology and Evolution. 2014

Title: *Identification of causal genomic loci in rats selected for tame and aggressive behavior*

Faraday Institute Essay Prize: 'Uses and Abuses of Biology':2nd Place 2012

Title: *Explore the ways in which contemporary genetics both challenges and underpins notions of human freedom, value and identity*

Presentations

Talk: Population Genetics CMPG Lab - Invited Speaker (Bern, Switzerland). 2015
Nice rats, nasty rats: The genetics of animal domestication

Talk: Walter M. Fitch Symposium, SMBE Meeting (Puerto Rico). 2014

Curriculum Vitae

Genetic variants underlying tameness and aggression

Talk: SMBE Meeting (Chicago, IL). 2013

Genetic variants underlying tameness and aggression

Talk: Rat Genomics and Models meeting (CSHL, NY). 2013

Genetic variants underlying tameness and aggression

Talk: Rat Genomics and Models meeting (Cambridge, UK). 2012

Genetic variants underlying tameness and aggression

Talk: Chromosome Conference (Novosibirsk, Russia). 2012

Genetic variants underlying tameness and aggression

Talk: Department of Anthropology - Invited Speaker (Yale, NH). 2011

Nice rats, nasty rats: Towards the genetics of animal domestication.

Talk: Rat Genomics and Models meeting (CSHL, NY). 2011

The genetics of animal domestication.

Talk: Rat Genomics and Models meeting (Kyoto, Japan). 2010

The genetic basis of tameness in a rat model.

Courses & Workshops

Phylogenetics Course

Leipzig (Germany), August 2015

Molecular Anthropology

Leipzig (Germany), Feb 2015

AWK training

Leipzig (Germany), April 2013

Unix training

Leipzig (Germany), February 2012

Linkage and Recombination in Genome sequences

Okinawa (Japan), May 2011

Introduction to Statistics

Leipzig (Germany), May 2011

Basics of Professional University Instruction/Teaching

Leipzig (Germany), April 2011

Voice and Speech Training, Rhetorics

Leipzig (Germany), March 2011

Introduction to Python

Leipzig (Germany), March 2011

Declaration of Independence

Herewith, I declare that I have conceived and written this thesis without any inadmissible help or material that has not been explicitly indicated. I have not previously attempted to complete this or another PhD thesis.

Leipzig, 1st of March, 2017

Author contribution statement, Alex Cagan

Thesis: Comparative genomic approaches to human evolutionary history

Author contribution statement:

Title: **Identification of genomic variants putatively targeted by selection during dog domestication**

Journal: BMC Evolutionary Biology

Authors: Alex Cagan, Torsten Blass

Part Alex Cagan:

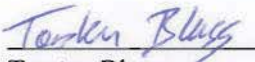
- Conceived the study
- Performed the analyses
- Wrote the manuscript

Part Torsten Blass:

- Contributed to statistical analyses



Alex Cagan



Torsten Blass

Author contribution statement, Alex Cagan

Thesis: Comparative genomic approaches to human evolutionary history

Author contribution statement:

Title: Genes and pathways selected during animal domestication

Journal: Submitted to eLife

Authors: Alex Cagan, Frank W. Albert, Irina Plyusnina[†], Lyudmila Trut, Gabriel Renaud, Frederic Romagné, Victor Wiebe, Rimma Kozhemjakina, Rimma Gulevich, Oleg Trapezov, Nikolay Yudin, Tatyana Alekhina, Ruslan Aitnazarov, Ludmila Trapezova, Yury Herbeck, Torsten Schöneberg, and Svante Pääbo

[†]Deceased

Part Alex Cagan:

- Designed the study
- Collected the rat and mink samples
- Collected the genomic data
- Performed the analyses
- Wrote the manuscript

Part Frank W. Albert

- Assisted in designing the study
- Contributed to the writing of the manuscript

Part Irina Plyusnina:

- Assisted in collecting the samples
- Maintenance of the rat lines
- Supervised the project

Part Lyudmila Trut:

- Assisted in collecting the samples
- Oversaw the experimental lines

Part Gabriel Renaud:

- Assisted in processing the data

Part Frederic Romagné:

- Assisted in selection analyses

Part Victor Wiebe:

- Prepared the rat and mink samples for sequencing

Part Rimma Kozhemjakina:

- Assisted in collecting the rat samples
- Maintenance of the rat lines

Part Rimma Gulevich:

- Assisted in collecting the rat samples
- Maintenance of the rat lines

Part Oleg Trapezov:

- Assisted in collecting the mink samples
- Maintenance of the mink lines

Part Nikolay Yudin:

- Assisted in collecting the mink samples
- Contributed to the writing of the manuscript

Part Tatyana Alekhina:

- Assisted in collecting the mink samples

Part Ruslan Aitnazarov:

- Assisted in collecting the mink samples

Part Ludmila Trapezova:

- Assisted in collecting the mink samples
- Maintenance of the mink lines

Part Yury Herbeck:

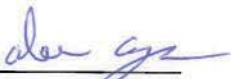
- Assisted in collecting the rat samples
- Contributed to the writing of the manuscript

Part Torsten Schöneberg:

- Assisted in collecting the rat samples
- Contributed to the writing of the manuscript

Part Svante Pääbo:

- Conceived the project
- Designed the study
- Supervised the project
- Wrote the manuscript



Alex Cagan

[Signature] _____
Frank W. Albert

[Signature] _____
Lyudmila Trut

[Signature] _____
Gabriel Renaud



Frederic Romagné



Victor Wiebe

[Signature] _____
Rimma Kozhemjakina

[Signature] _____
Rimma Gulevich

[Signature] _____
Oleg Trapezov

[Signature] _____
Nikolay Yudin

[Signature] _____
Alex Cagan

February 20, 2017
Frank W. Albert



[Signature] _____
Lyudmila Trut

[Signature] _____
Gabriel Renaud

[Signature] _____
Frederic Romagné

[Signature] _____
Victor Wiebe

[Signature] _____
Rimma Kozhemjakina

[Signature] _____
Rimma Gulevich


[Signature] _____
Oleg Trapezov

[Signature] _____
Nikolay Yudin

[Signature] _____
Tatyana Alekhina

[Signature] _____
Alex Cagan


[Signature] _____
Frank W. Albert

[Signature]  _____
Lyudmila Trut

[Signature] _____
Gabriel Renaud

[Signature] _____
Frederic Romagné

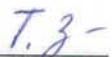
[Signature] _____
Victor Wiebe

[Signature]  _____
Rimma Kozhemjakina

[Signature]  _____
Rimma Gulevich

[Signature] _____
Oleg Trapezov

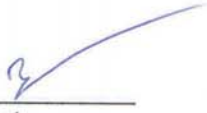
[Signature] _____
Nikolay Yudin

[Signature]  _____

Tatyana Alekhina

[Signature] _____
Ruslan Aitnazarov

[Signature] _____
Ludmila Trapezova

[Signature]  _____
Yury Herbeck

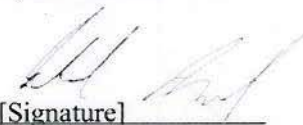
[Signature] _____
Torsten Schöneberg

[Signature] _____
Svante Pääbo

[Signature] _____
Alex Cagan

[Signature] _____
Frank W. Albert

[Signature] _____
Lyudmila Trut


[Signature] _____
Gabriel Renaud

[Signature] _____
Frederic Romagné

[Signature] _____
Victor Wiebe

[Signature] _____
Rimma Kozhemjakina

[Signature] _____
Rimma Gulevich

[Signature] _____
Oleg Trapezov

[Signature] _____
Nikolay Yudin

[Signature]

Victor Wiebe

[Signature]

Rimma Kozhemjakina

[Signature]

Rimma Gulevich



Oleg Trapezov

[Signature]

Nikolay Yudin

[Signature]

Tatyana Alekhina

[Signature]

Ruslan Aitnazarov



Ludmila Trapezova

[Signature]

Yury Herbeck

[Signature]

Torsten Schöneberg

[Signature]

Svante Pääbo

[Signature] _____
Alex Cagan

[Signature] _____
Frank W. Albert

[Signature] _____
Lyudmila Trut

[Signature] _____
Gabriel Renaud

[Signature] _____
Frederic Romagné

[Signature] _____
Victor Wiebe


[Signature] _____
Rimma Kozhemjakina

[Signature] _____
Rimma Gulevich

[Signature] _____
Oleg Trapezov

[Signature] *N. Yudin*
Nikolay Yudin

[Signature] _____
Tatyana Alekhina

[Signature] 
Ruslan Aitnazarov

[Signature] _____
Ludmila Trapezova

[Signature] _____
Yury Herbeck

[Signature] _____
Torsten Schöneberg

[Signature] _____
Svante Pääbo

Nikolay Yudin

[Signature] _____

Tatyana Alekhina

[Signature] _____

Ruslan Aitnazarov

[Signature] _____

Ludmila Trapezova

[Signature] _____

Yury Herbeck



[Signature] _____

Torsten Schöneberg

Svante Pääbo


[Signature] _____
Tatyana Alekhina

[Signature] _____
Ruslan Aitnazarov

[Signature] _____
Ludmila Trapezova

[Signature] _____
Yury Herbeck

[Signature] _____
Torsten Schöneberg



Svante Pääbo

Author contribution statement, Alex Cagan

Thesis: Comparative genomic approaches to human evolutionary history

Author contribution statement:

Title: Natural Selection in the Great Apes

Journal: Molecular Biology and Evolution

Authors: Alex Cagan, Christoph Theunert, Hafid Laayouni, Gabriel Santpere, Marc Pybus, Ferran Casals, Kay Prüfer, Arcadi Navarro, Tomas Marques-Bonet, Jaume Bertranpetit, and Aida M Andrés

Part Alex Cagan:

- Designed the study
- Performed the gene overlap analyses
- Performed the HKA analyses and ELS analyses
- Performed the gene overlap analyses
- Performed the gene ontology analyses
- Wrote the first draft of the manuscript

Part Christoph Theunert:

- Performed the HKA and ELS analyses
- Contributed to the writing of the manuscript

Part Hafid Laayouni:

- Performed the FWH analyses
- Contributed to the writing of the manuscript

Part Gabriel Santpere:

- Filtered the data
- Performed the MK and alpha analyses
- Contributed to the writing of the manuscript

Part Marc Pybus:

- Created the selection browser
- Contributed to genomic analyses

Part Ferran Casals:

- Created the selection browser
- Contributed to genomic analyses

Part Kay Prüfer:

- Contributed to the ELS analyses

Part Arcadi Navarro:

- Conceived the project
- Contributed to the writing of the manuscript

Part Tomas Marques-Bonet:

- Provided necessary data and information
- Assisted with processing the data

- Contributed to the writing on the manuscript

Part Jaume Bertranpetit:

- Conceived the project
- Designed the study
- Supervised the project
- Wrote the manuscript

Part Aida M Andrés:

- Conceived the project
- Designed the study
- Supervised the project
- Wrote the manuscript



Alex Cagan

[Signature] _____

Christoph Theunert

[Signature] _____

Hafid Laayouni

[Signature] _____

Gabriel Santpere

[Signature] _____

Marc Pybus

[Signature] _____

Ferran Casals



Kay Prüfer

[Signature] _____


Arcadi Navarro

[Signature] _____

Tomas Marques-Bonet

[Signature] _____

Jaume Bertranpetit



Aida M Andrés

[Signature] _____
Alex Cagan



Christoph Theunert

[Signature] _____
Hafid Laayouni

[Signature] _____
Gabriel Santpere

[Signature] _____
Marc Pybus

[Signature] _____
Ferran Casals

[Signature] _____
Kay Prüfer

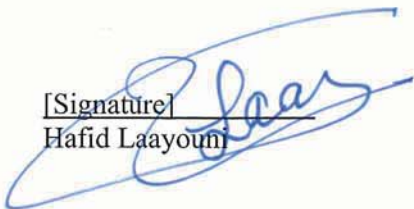
[Signature] _____
Arcadi Navarro

[Signature] _____
Tomas Marques-Bonet

[Signature] _____
Jaume Bertranpetit


[Signature] _____
Alex Cagan

[Signature] _____
Christoph Theunert

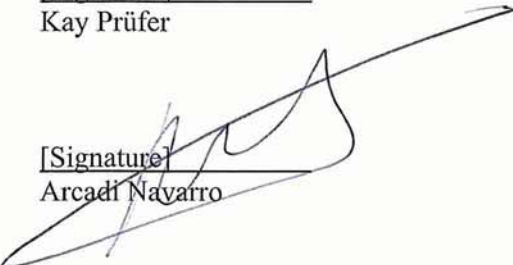
[Signature]  _____
Hafid Laayouni

[Signature] _____
Gabriel Santpere

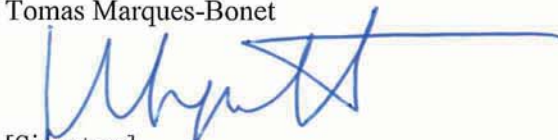
[Signature] _____
Marc Pybus


[Signature] _____
Ferran Casals

[Signature] _____
Kay Prüfer

[Signature]  _____
Arcadi Navarro

[Signature] _____
Tomas Marques-Bonet


[Signature] _____
Jaume Bertranpetit


[Signature] _____
Aida M Andrés

Alex Cagan

[Signature]
Christoph Theunert

[Signature]
Hafid Laayouni

[Signature]
Gabriel Santpere



[Signature]
Marc Pybus

[Signature]
Ferran Casals

Kay Prüfer

[Signature]
Arcadi Navarro

[Signature]
Tomas Marques-Bonet

[Signature]
Jaume Bertranpetit

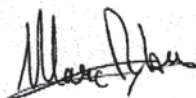
Aida M Andrés

[Signature]
Alex Cagan

[Signature]
Christoph Theunert

[Signature]
Hafid Laayouni

[Signature]
Gabriel Santpere


[Signature]
Marc Pybus

[Signature]
Ferran Casals

[Signature]
Kay Prüfer

[Signature]
Arcadi Navarro

[Signature]
Tomas Marques-Bonet

[Signature]
Jaume Bertranpetit

[Signature]
Aida M Andrés

[Signature] _____
Alex Cagan

[Signature] _____
Christoph Theunert

[Signature] _____
Hafid Laayouni

[Signature] _____
Gabriel Santpere

[Signature] _____
Marc Pybus

[Signature] _____
Ferran Casals

[Signature] _____
Kay Prüfer

[Signature] _____
Arcadi Navarro



[Signature] _____
Tomas Marques-Bonet

[Signature] _____
Jaume Bertranpetit