Long-term balancing selection in the genomes of humans and other great apes

Von der Fakultät für Biowissenschaften, Pharmazie und Psychologie

der Universität Leipzig

genehmigte

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium

Dr. rer. nat.

vorgelegt

von

Master of Science João Teixeira

geboren am 30.07.1987 in Porto, Portugal

Dekan:	Prof. Dr. Tilo Pompe
Gutachter:	Prof. Dr. Svante Pääbo
	Prof. Dr. Tomas Marques-Bonet

Tag der Verteidigung: 02.09.2016

Long-term balancing selection in the genomes of humans and other great apes

Der Fakultät für Biowissenschaften, Pharmazie und Psychologie

der Universität Leipzig

eingereichte

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium

Dr. rer. nat.

vorgelegt

von Master of Science João Teixeira

geboren am 30.07.1987 in Porto, Portugal

Leipzig, den 27.04.2016

BIBLIOGRAPHISCHE DARSTELLUNG

João Teixeira

Long-term balancing selection in the genomes of humans and other great apes

Fakultät für Biowissenschaften, Pharmazie und Psychologie

Universität Leipzig

Dissertation

161 Seiten, 130 Literaturangaben, 20 Abbildungen, 19 Tabellen

Abstract

Balancing selection maintains advantageous genetic diversity in populations through a variety of mechanisms including overdominance, negative frequency-dependent selection, temporal or spatial variation in selective pressures, and pleiotropy. If environmental pressures are constant through time, balancing selection can affect the evolution of selected loci for millions of years, and its targets might be shared by different species. This thesis is comprised of two different approaches aimed at detecting shared signatures of balancing selection in the genomes of humans and other great apes.

In the first part of the thesis, we focus on extreme loci where the action of balancing selection has maintained several coding trans-species polymorphisms in humans, chimpanzees and bonobos. These trSNPs segregate since the common ancestor of the *Homo-Pan* clade and have survived for ~14 million years of independent evolution. These loci show the characteristic signatures of long-term balancing selection, as they define haplotypes with high genetic diversity that show cluster of sequences by allele rather than by species, and segregate at intermediate allele frequencies. Apart from several trSNPs in the MHC region, we were able to uncover a non-synonymous trSNP in the autoimmune gene *LAD1*.

In the second part of the thesis we explore shared signatures of balancing selection outside trSNPs. We first implement a genome scan designed to detect signatures of balancing selection using NCD2 in the genomes of nine great ape species, including chimpanzee, bonobo, gorilla and orangutan. We show that targets of balancing selection are shared between species that have diverged millions of years ago, and that this observation cannot be explained by shared ancestry. We further demonstrate that targets of balancing selection primarily affect the evolution of genic regions of the genome, although we see evidence for their involvement in the regulation of gene expression.

Overall, we provide comprehensive evidence that similar environmental pressures maintain advantageous diversity through the action of balancing selection in humans and other great apes, notwithstanding the deep divergence times between many of these species.

For my parents Helena and Raúl.

Não sou nada.
Nunca serei nada.
Não posso querer ser nada.
À parte isso, tenho em mim todos os sonhos do mundo.

Fernando Pessoa *in Tabacaria*

Table of contents

1. Thesis summary

The vast majority of genetic variants in a population are effectively neutral and, therefore, the frequency of their alleles varies as a function of genetic drift (Kimura 1983). Nevertheless, all biological populations harbor mutations that impact the fitness of individuals and their chance to reproduce. Such mutations are then targeted by natural selection and different alleles can be either advantageous and increase in frequency through positive selection (e.g. Sabeti et al. 2006), or deleterious and eliminated from the population by purifying selection (e.g. Ward and Kellis 2012). In other cases, however, genetic diversity is on itself advantageous and alleles are maintained in populations by balancing selection (e.g. Andrés et al. 2009).

A classic example of balancing selection in humans affects the evolution of the *β*-globin gene in populations where malaria is endemic. In this locus, homozygous individuals for the wild-type allele (HbA/HbA genotype) are healthy but may be infected by *Plasmodium falciparum* and develop malaria, whereas homozygous for the mutant allele (HbS/HbS genotype) suffer from sickle-cell anemia and have a highly reduced life expectancy. Heterozygous individuals, on the other hand, are healthy and more resistant to malaria than HbA homozygotes, whereby the sickle-cell variant is maintained in these populations due to heterozygote advantage (Allison 1956; Pasvol et al. 1978). There are other mechanisms through which balancing selection maintains advantageous genetic diversity in populations including negative frequency-dependent selection, temporal or spatial variation in selective pressures, and pleiotropy (Wright S. 1939; Pasvol et al. 1978; Gillespie 1978; Gigord et al. 2001; Muelenbachs et al. 2008; Gendzekhadze et al. 2009; Hughes et al. 2013).

The study of balancing selection in humans has mainly been driven by candidate gene approaches (Hughes and Nei 1989; Prugnolle et al. 2005; Fumagalli et al. 2009; Wooding et al. 2005), and provided but a limited comprehension on how it affects the evolution of the human genome. While knowledge of balancing selection in humans has certainly improved with recent implementations of genome-wide scans (Andrés et al. 2009; Rasmussen et al. 2014; DeGiorgio et al. 2014), no study has yet attempted to unveil its importance in adaptation among our closest living relatives, the great apes. This would allow, for example, to understand how targets of balancing selection are conserved between closely related species and bring to light loci on which selective pressures have changed through time.

This is particularly relevant because selective pressures can affect the evolution of a particular locus for millions of years and, therefore, instances of balancing selection might be *shared* between species. In fact, it is even theoretically possible for balanced polymorphisms to survive the differentiation and split of populations into different species, whereby some single nucleotide polymorphisms (SNPs) will segregate in different species, resulting in trans-species polymorphisms – trSNPs (Muirhead et al. 2002; Asthana et al. 2005; Charlesworth 2006; Cagliani et al. 2010; Cagliani et al. 2012; Ségurél et al. 2012; Leffler et al. 2013; Key et al. 2014). In species with old divergence, where trSNPs are not expected under neutrality, they are a hallmark of long-term balancing selection and highlight extreme examples of conservation of selective pressures (Asthana et al. 2005; Ségurél et al. 2013; Leffler et al. 2013). Uncovering trSNPs in the genome should therefore help bring to light loci on which genetic diversity is advantageous in the genome.

Probably the most renowned example of a locus evolving under long-term balancing selection and containing known trSNPs is the major-histocompatibility (MHC)

cluster. This region has a key role in antigen presentation and is therefore highly relevant for the action of the immune system, and trSNPs have been uncovered in a variety of different vertebrate clades, from fish (Graser et al. 1996) and birds (Kikkawa et al. 2009; Sutton et al. 2013), to rodents (Cutrera et al. 2007) and primates (Klein et al. 1993; Asthana et al. 2005; Loisel et al. 2006; Ségurél et al. 2012; Leffler et al. 2013). Notwithstanding the relevance of the MHC cluster, only a few studies have addressed the existence of trSNPs in primates that are outside the region, all focusing on pre-determined, specific genes (Cagliani et al. 2010; Cagliani et al. 2012; Ségurél et al. 2012). Recently, a genome-wide study identified a few trans-species haplotypes in humans and chimpanzees, albeit none of these contained protein-coding SNPs (Leffler et al. 2013), which seems surprising given that natural selection is known to preferentially affect the evolution of coding regions of the genome. Nevertheless, this study focused on the existence of at least two trSNPs in humans and chimpanzees that are in complete linkage, whereby cases where balancing selection maintains a single trSNP in both species were not considered.

The first part of this thesis focuses on understanding the extent to which balancing selection has been conserved among humans and close living relatives by focusing on uncovering trSNPs in a set composed by the complete exomes of 20 humans (*Homo sapiens sapiens*), 20 central chimpanzees (*Pan troglodytes troglodytes*) and 20 bonobos (*Pan paniscus*). We start by estimating the probability of observing a trSNP in the three species under neutrality by implementing a model based on coalescent theory and using neutral simulations on realistic demographic scenarios for the human, chimpanzee and bonobo populations (Prado-Martínez et al. 2013). We show this probability to be very low (4.0×10^{-10}) and that given the number of single

nucleotide polymorphisms (SNPs) uncovered in our set of 20 human individuals, the expected number of trSNPs with both chimpanzees and bonobos is virtually zero (5.0 x 10^{-5}). Nevertheless, and after eliminating SNPs likely arising from genotype sequencing errors and recurrent mutations, we were able to find a total of 8 coding trSNPs (of which 5 are non-synonymous) that have strong evidence of being maintained by balancing selection for over 15 million years of independent evolution. These trSNPs represent hallmarks of the action of balancing selection in shaping the genomes of these three species and potentially affect protein structure and function. All trSNPs exhibit the typical signatures of balancing selection, such as defining a haplotype that shows clustering of sequences by allele rather than by species, segregating at intermediate frequencies in all three species, and lying in a locus with unusually high levels of genetic polymorphism. Interestingly, apart from previously known trSNPs in the MHC region, we found one non-synonymous trSNP (Leucine->Proline) segregating in the gene *Ladinin 1*(*LAD1)*, which encodes an anchoring filament protein that maintains cohesion at the dermal-epidermal junction, and is an autoantigen associated with linear IgA disease, an autoimmune condition that causes blistering of the skin (Ishiko et al. 1996; Marinkovich et al. 1996; Motoki et al. 1997; McKee et al. 2005). Apart from resulting in a different protein sequence, the two alleles found at the trSNP in *LAD1* are associated with differences in gene expression, which opens the possibility for this trSNP to have, additionally, regulatory effects. The biological basis for long-term balancing selection acting on *LAD1* remains elusive. Nevertheless, genes associated with cell-adhesion have been previously proposed as targets of balancing selection (Andrés et al. 2009; Fumagalli et al. 2009; 2011). Furthermore, balancing selection has also been proposed to affect the evolution of autoimmune genes given that inflammatory response must be efficient and, at the

same time, moderate in order to avoid self-recognition by the immune system (Ferrer-Admetlla et al. 2008). Notwithstanding, it is possible that autoimmune reactions are influenced by genetic diversity maintained by balancing selection.

The first part of this work provides evidence, for the first time, of the existence of trSNPs between humans and both chimpanzees and bonobos outside of the MHC cluster. This trSNP highlights a region of the genome where balancing selection has been acting since at least the common ancestor of the three species, representing ~14 million years of independent evolution.

Thus, the loci containing trSNPs between humans and other primates unveil only some of the most extreme examples of long-term balancing selection, and represent a limited view of its targets in the genome. As a consequence, they may not represent well the degree of *sharing* of selective pressures across species that maintain advantageous diversity in populations.

A new dataset containing genome-wide sequencing for all major clades of non-human great apes has become recently available (Prado-Martínez et al. 2013), which allows for investigating for the first time how balancing selection affected the evolution of the genome of humans' closest living relatives. Moreover, such analysis makes it possible to better understand whether targets of selection have been conserved across species that diverged several million years in the past outside known examples provided by trSNPs. In the second part of this work, I describe a strategy implemented to identify targets of balancing selection in the genomes of 9 different great ape species: 4 common chimpanzee subspecies (*Pan troglodytes*) including western (*P. t. verus*), eastern (*P. t. schweinfurthii*), central (*P. t. troglodytes*) and Nigeria-Cameroon (*P. t. ellioti*) chimpanzees; bonobos (*Pan paniscus*); 2 gorilla

subspecies including western (*Gorilla gorilla gorilla*) and eastern (*G. beringei graueri*) lowland gorillas; 2 subspecies of orangutans from Sumatra (*Pongo abelli*) and Borneo (*P. pygmaeus*). We aim to identify targets of balancing selection using a novel statistic (Non-Central Deviation, or NCD; Bitarello et al. in prep), which combines information on the number of polymorphisms and the frequency of their alleles in a given genomic region, and that has high power to identify loci with the patterns of genetic diversity expected under balancing selection (Bitarello et al. in prep). First, we run simulations with and without balancing selection under realistic demographic models for each great ape species (Prado-Martínez et al. 2013) to show that NCD has high power to detect long-term balancing selection also in the great apes. We then calculate NCD using a sliding-window approach across the genomes of 9 great ape species using human as an outgroup. We focus on windows showing the lowest NCD_{value} in each species, which are enriched for targets of long-term balancing selection. In fact, we demonstrate that these windows include well known targets of balancing selection (e.g MHC genes) and, performing a Gene Ontology analysis, reveal that they are enriched in categories involved in immune response, particularly antigen presentation pathway (HLA genes in the MHC cluster). Moreover, these sets of windows overlap with protein coding regions of the genome more often than expected by chance and, at the same time, often overlap with regulatory regions. Together, these results indicate that the tails of the NCD distribution are, in fact, enriched for true targets of balancing selection.

We then investigate whether candidate targets of balancing selection are shared by different species by comparing the proportion of candidate windows shared by pairs of species. Remarkably, we observe higher sharing among the candidate targets of balancing selection than expected given the background correlation in patterns of

polymorphism as a consequence of shared ancestry. This is observed across all pairwise comparisons, even among species that diverged millions of years ago, like chimpanzees and gorillas or orangutans. The amount of sharing is indeed higher than expected given shared ancestry between species, as shown by comparisons with the remainder of the genome and by using neutral simulations of great ape demographic history.

Taken together, these results strongly suggest that targets of balancing selection are shared between species even when divergence is millions of years old, which indicates that selective pressures acting on the genome of these species are, to some extent, similar. Moreover, it is likely that some of these windows include targets where balancing selection is acting since the common ancestor of different species. Among the top candidates are 5 genes that include at least one candidate window in all 9 great ape species analyzed in our study: *HLA-DRB1*, *HLA-DRB5*, *HLA-DQB1*, *ATN1* and *GARNL4*. The three HLA genes, located in the MHC region, are also candidates for balancing selection in humans (Bitarello et al. in prep).

Overall, this thesis comprises evidence for the influence of long-term balancing selection in the genomes of humans and closest relatives. We show that targets of balancing selection often overlap genic regions of the genome, and that a high proportion of candidate genes are associated with immune response. We demonstrate that a higher than expected proportion of targets is shared among species and thus evolving under similar selective pressures. Finally, we provide the first catalog of candidate targets of balancing selection in the genomes of non-human great apes.

The findings presented in the first part of this work resulted in the publication of the manuscript **"*Long-term balancing selection in LAD1 maintains a missense trans-***

***species polymorphism in humans, chimpanzees and bonobos*"** in the journal Molecular Biology and Evolution (Teixeira et al. MBE 2015). A manuscript describing the second part of the findings reported in this work is currently under preparation.

2. Zusammenfassung

Die überwiegende Mehrheit der genetischen Variation in einer Population ist neutral und damit ist ihrer Allelfrequenz abhängig vom genetischen Drift (Kimura 1983). Dennoch tragen alle Populationen Mutationen, die sich auf die Fitness von Individuen auswirken und damit ihre Chance sich zu reproduzieren beeinflussen. Die Frequenz solcher vorteilhaften Mutationen kann dann durch gezielte natürliche Selektion ansteigen (zum Beispiel Sabeti et al., 2006), bzw. beeinträchtigende Mutationen werden aus der Population eliminiert (z.B. Ward und Kellis 2012). In besonderen Fällen jedoch ist die genetische Vielfalt an sich vorteilhaft und verschiedene Allele werden in der Bevölkerung aufrechterhalten durch sogenannte *balancing selection* (z.B. Andrés et al. 2009).

Ein klassisches Beispiel für balancing selection beim Menschen ist die Entwicklung des β-Globin-Gens in Populationen die in Malariagebieten leben. Dort sind Individuen die homozygot für das Wildtyp-Allel (HbA / HbA Genotyp) sind zwar gesund, aber gefährdet sich mit Plasmodium falciparum zu infizieren, während homozygote Individuen für das mutierte Allel (HBS / HbS Genotyp) unter Sichelzellenanämie leiden und eine stark reduzierte Lebenserwartung haben. Heterozygote Individuen andererseits sind gesund und resistent gegen Malaria und haben keine Sichelzellanämie. Dadurch wird der heterozygote Genotyp in der Population aufgrund von einem sog. heterozygoten Vorteil gehalten (Allison 1956; Pasvol et al. 1978). Es gibt auch andere Mechanismen, wodurch balancing selection vorteilhafte genetische Vielfalt in der Bevölkerung erhält, z. B. negative frequenzabhängige Selektion, zeitliche oder räumliche Variation im selektiven Druck

und Pleiotropie (Wright S. 1939; Pasvol et al. 1978; Gillespie 1978; Gigord et al. 2001; Muehlenbachs et al. 2008; Gendzekhadze et al. 2009; Hughes et al. 2013).

Lange Zeit wurde balancing selection beim Menschen in erster Linie durch Kandidaten-Gene beschrieben (Hughes und Nei 1989; Prugnolle et al. 2005; Fumagalli et al. 2009; Bamshad et al. 2009; Wooding et al. 2005). Dadurch haben wir nur ein begrenztes Verständnis darüber, wie wichtig balancing selection bei der Entwicklung des menschlichen Genoms war und ist. Durch die zusätzliche Analyse von genomweiten Scans hat sich unser Verständnis über balancing selection verbessert (Andrés et al. 2009; Rasmussen et al. 2014; DeGiorgio et al. 2014). Bisher hat keine Studie versucht die Bedeutung von balancing selection bei der Anpassung der uns nächsten lebenden Verwandten, den Menschenaffen, zu enthüllen. Dies würde es ermöglichen, zum Beispiel, zu verstehen, wie Zieleregionen der balancing selection zwischen eng verwandten Spezies konserviert wurden bzw. wie sich der Selektionsdruck durch die Zeit verändert hat.

Dies ist besonders relevant, da balancing selection die Entwicklung eines bestimmten Locus für Millionen von Jahren beeinflussen kann und damit Regionen die unter balancing selection stehen zwischen Arten geteilt werden kann. In der Tat ist es auch theoretisch möglich, dass Mutationen unter balancing selection über die Differenzierung und Spaltung von Gruppen in verschiedene Arten hinaus erhalten bleiben, und damit zu sogenannten trans-Spezies Polymorphismen -. trSNPs werden (Muirhead et al 2002; Asthana et al . 2005; Charlesworth 2006; Cagliani et al. 2010; Cagliani et al. 2012; Ségurel et al. 2012; Leffler et al. 2013; Key et al. 2014). Bei Arten mit alter Divergenz, wo trSNPs nicht unter Neutralität zu erwarten sind, sind sie ein Zeichen von langfristiger balancing selection und ein extremes Beispiele für die

Erhaltung von Mutationen (Asthana et al. 2005; Ségurel et al. 2013; Leffler et al. 2013). trSNPs im Genom zu erkennen sollte daher dazu beitragen, loci im Genom zu erkennen wo genetische Vielfalt von besonderem Vorteil war und ist.

Wahrscheinlich das bekannteste Beispiel für langfristige balancing selection und trSNPs ist der Major-Histocompatibility (MHC) Cluster. Diese Region hat eine wichtige Rolle bei der Antigenpräsentation und ist daher in hohem Maße relevant für ein funktionierendes Immunsystems. In dieser Region sind in einer Vielzahl von verschiedenen Vertebraten wie Fisch (Graser et al. 1996), Vögel (Kikkawa et al 2009; Sutton et al. 2013), Nagetiere (Cutrera et al. 2007) und Primaten (Klein et al.; 1993, Asthana et al. 2005; Loisel et al. 2006; Ségurel et al. 2012; Leffler et al. 2013) trSNPs beschrieben. Ungeachtet der Bedeutung des MHC clusters, gibt es nur wenige Studien die die Existenz von trSNPs in Primaten berichten. Wobei die meißten dieser Studien nur gezielte Gene untersucht haben (Cagliani et al. 2010; Cagliani et al. 2012; Segurel et al. 2012). Vor kurzem wurde in einer genomweiten Studie neue trans Spezies Haplotypen in Menschen und Schimpansen analysiert, wenn auch keine der neuen Regionen Protein-codierenden SNPs enthalten (Leffler et al. 2013). Dies ist überraschend da natürliche Selektion bevorzugt die Evolution von codierenden Regionen des Genoms beeinflusst. Weiterhin konzentrierte sich diese Studie auf Regionen von mindestens zwei trSNPs die in Menschen und Schimpansen vollständig verknüpft sind, während die Möglichkeit einzelner trSNP in beiden Spezies nicht berücksichtigt wurde.

Der erste Teil dieser Arbeit konzentriert sich das Ausmaß und das Verständnis, von balancing selection und die Aufdeckung von trSNPs zwischen Menschen und ihren

nächsten, lebenden Verwandten. Dafür nutzen wir vollständige exomes von 20 Menschen (*Homo sapiens sapiens*), 20 Schimpansen (*Pan troglodytes troglodytes*), und 20 Bonobos (*Pan paniscus*). Wir beginnen damit die Wahrscheinlichkeit zu schätzen, einen trSNP in den drei Spezies unter Neutralität zu beobachten. Dafür entwickeln wir ein Modell basierend auf der coalescent Theorie und nutzen neutrale Simulationen mit realistischen demographischen Szenarien für Menschen, Schimpansen und Bonobos (Prado-Martinez et al. 2013). Wir zeigen dass die Wahrscheinlichkeit sehr gering ist ($4{,}0 \times 10^{-10}$) und dass bei der Anzahl von Mutationen (SNPs) in unserem Set, die erwartete Anzahl von trSNPs mit beiden Spezies (Schimpansen und Bonobos) praktisch null ist ($5{,}0 \times 10^{-5}$). Nachdem wir SNPs die wahrscheinlich aufgrund von Genotyp Sequenzierungsfehler und wiederkehrenden Mutationen im Datensatz sind eliminiert wurden finden wir 8 kodierende trSNPs (von denen sind 5 non-synonymous). Ausserdem tragen die loci starke Signaturen dafür dass diese SNPs 15 Millionen Jahre lang durch balancing selection beibehalten wurden trotz der unabhängigen Entwicklung der drei Spezies. Diese trSNPs zeigen Markenzeichen von balancing selection in den Genomen dieser drei Arten, die möglicherweise Protein-Struktur und Protein-Funktion beeinflussen. Alle trSNPs zeigen typische Signaturen von balancing selection, beispielsweise, indem die Haplotypen in einem phylogenetischen Baum zusammenfallen anstelle der Arten, des weiteren liegen die trSNPs in intermediate Frequenzen in allen drei Arten vor, und die genetischen loci weisen besonders ungewöhnliche, hohe, genetischen Variabilität auf. Interessanterweise, abgesehen von bisher bekannten trSNPs in der MHC-Region, fanden wir einen non-synonymous trSNP (leucin> Prolin) in dem Gen Ladinin 1 (*LAD1*), das ein Protein für Verankerungsfilament kodiert, wodurch der Zusammenhalt an der dermal-epidermalen Verbindung aufrechterhalten wird. Des

weiteren ist es ein Autoantigen und mit der linearen IgA Krankheit verbunden, eine Autoimmunerkrankung (Ishiko et al 1996; Marinkovich et al 1996; Motoki et al 1997; McKee et al 2005). Außer dass beide Allele unterschiedliche Proteinsequenzen ergeben, sind die beiden Allele des trSNPs in *LAD1* mit Unterschieden in der Genexpression assoziiert, was die Möglichkeit eröffnet diese trSNP haben zusätzlich regulatorische Effekte. Dennoch, die biologische Basis für die langfristige balancing selection auf *LAD1* bleibt schwer zu fassen. Allerdings wurden Gene die im Zusammenhang mit Zell-Adhäsion stehen schon zuvor als Ziele von balancing selection vermutet (Andrés et al. 2009; Fumagalli et al. 2009, 2011). Weiterhin scheint balancing selection die evolution von Autoimmungenen zu beeinflussen, gemessen daran dass die Entzündungsreaktion effizient sein muss und gleichzeitig schwach um Selbst-Erkennung durch das Immunsystem zu verhindern (Ferrer-Admetlla et al. 2008). Abweichend ist es möglich, dass Autoimmunreaktionen durch genetische Diversität ausgelöst werden und evtl. durch balancing selection begünstigt werden. Der erste Teil dieser Arbeit belegt die Existenz von trSNPs zwischen Menschen und sowohl Schimpansen als auch Bonobos außerhalb des MHC-Clusters. Diese trSNP beschreit eine Region des Genoms, wo balancing selection seit dem gemeinsamen Vorfahren der drei Spezies vor ~ 14.000.000 Jahren wirkt.

Genomische Regionen mit trSNPs zwischen Menschen und anderen Primaten umfassen nur einen kleinen Teil der extremsten Beispiele für Langzeit-Balancing-Selection und ermöglichen damit nur einen limitierten Blick auf Kandidaten für Balancing-Selection. Auch deshalb sind trSNPs unter Umständen nicht representativ um den gemeinsamen Selektionsdruck zwischen Spezies zu reflektieren.

Ein kürzlich veröffentlicher Datensatz mit Genomdaten für alle Menschenaffen ermöglicht nun Balacing Selection in Menschen und seinen nächsten lebenden Verwandten zu untersuchen. Besonders die Frage, ob Kandidaten zwischen diesen Spezies konserviert sind, ermöglicht ein besseres Verständnis über die Rolle von Balancing Selection, die über das vorhandene Wissen, erlangt durch trSNPs, hinaus geht. Im zweiten Teil meiner Arbeit beschäftigt sich mit der Suche nach Kandidaten für Balancing Selection in 9 Menschenaffen : 4 Schimpansen sub-Spezies (*Pan troglodytes*): western (*P. t. verus*), eastern (*P. t. schweinfurthii*), central (*P. t. troglodytes*) und Nigeria-Cameroon (*P. t. ellioti*) Schimpansen; Bonobos (*Pan paniscus*); 2 gorilla sub-Spezies: western (*Gorilla gorilla gorilla*) und eastern (*G. beringei graueri*) Flachlandgorillas; 2 sub-Spezies Orang-Utans aus Sumatra (*Pongo abelli*) und Borneo (*P. pygmaeus*).

Zur Identifizierung von Balancing-Selection-Kandidaten nutzen wir eine Statistik (Non-Central Deviation, or NCD; Bitarello et al. in prep), die Informationen über Polymorphismen und deren Allelfrequenz kombiniert um mit hoher Konfidenz Regionen mit erhöhter Diversität zu finden, die ein Hauptmerkmal von Balancing Selection ist (Bitarello et al. in prep). Um die Eignung von NCD Kandidaten für Balancing Selection zu finden, haben wir zunächst Simulationen mit realistischen demografischen Parametern für alle Menschenaffen (Prado-Martínez et al. 2013) durchgeführt. Im nächsten Schritt haben wir NCD in genomischen Fenstern für alle 9 Menschenaffen berechnet. Fenster mit den geringsten NCD-Werten in jeder Spezies wurden als kandidaten definiert, da diese für Kandidaten für Balancing Selection angereichert sein sollten.

In der Tat sind wir in der Lage zu zeigen, dass Fenster mit niedrigen NCD-Werten bekannte Kandidaten (MHC Gene) enthalten. Wir fanden auch eine

Überrepräsentation von Kategorien assoziiert mit Immunabwehr in der Gene Ontology für Gene in Fenstern mit niedrigen NCD-Werten, im Speziellen Kategorien, die Antigen-Pathways repräsentieren (HLA gene im MHC). NCD-Fenster mit niedrigen Werten überlappen signifikant häufig mit Protein-codierenden Genen und genomischen Regionen, die regulatorische Funktionen besitzen. Diese Resultate bestärken, dass Regionen in Fenstern mit niedrigen NCD Werten in der Tat angereichert sind mit echten Kanditaten für Balancing Selection.

Darüber hinaus haben wir den Überlapp von Kandiatenfenstern zwischen verschiedenen Spezies getestet und fanden das diese für alle paarweisen Speziesvergleiche, und damit auch für Spezies die Millionen von Jahren divergierten, stärker überlappten als erwartet im Vergleich mit neutralen Simulationen basierend auf dem demografischen Hintergrund der Menschenaffen.

Die Ergebnisse aus beiden Analysen zeigen, dass die Kandidaten für Balancing Selection selbst zwischen Spezies, die vor mehreren Millionen von Jahren divergiert sind, öfter zusammen zu finden sind und damit ähnlichen Selektiondruck ausgesetzt zu sein scheinen. Es ist ausserdem wahrscheinlich, dass die Kandidatenfenster genomische Regionen enthalten, die dem Selektionsdruckdruck schon seit dem gemeinsamen Vorfahren all der hier untersuchten Primatenspezies ausgesetzt sind. Unter den ersten 5 geteilten Genen in allen 9 Menschenaffen sind: *HLA-DRB1*, *HLA-DRB5*, *HLA-DQB1*, *ATN1* and *GARNL4*. Die 3 HLA-Gene innerhalb des MHC sind auch Kandidaten für Balancing Selection im Menschen (Bitarello et al. in prep).

Zusammengenommen zeigen die Ergebnisse dieser Arbeit den Einfluss von Balancing Selection auf die Genome von Menschen und deren nächsten Verwandten. Wir zeigen, dass Kandidaten für Balancing Selection öfter Gene überlappen und mit

Immunprozessen assoziiert sind. Ausserdem sind diese Kandidaten öfter gemeinsam in mehreren Spezies zu finden und damit unter gleichen Selektionsdruck. Mit diesen Ergbnissen stellen wir den ersten Katalog für Kandidaten für Balancing Selection in Menschenaffen zusammen.

Die Ergebnisse der ersten Abschnitts sind veröffentlicht in **"*Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees and bonobos*"** im Wissenschaftsjournal Molecular Biology and Evolution (Teixeira et al. MBE 2015). Ein Manuskript über die Ergebnisse des zweiten Teils wird zur Zeit zusammengestellt.

3. Introduction

3.1 The forces shaping evolution

The evolution of life on Earth is governed by the interplay of different evolutionary processes, responsible for the generation, elimination and maintenance of genetic diversity (Nei 1975). These processes are behind the continuous differentiation of populations and species through time since the emergence of life, a several billion-years process that drives the observable complexity, beauty and dissimilitude of living organisms.

Evolutionary processes can be regarded as adaptive and non-adaptive, depending on their effects on the fitness of individuals (Lynch 2007). Arguably the most known of such processes is *natural selection*, proposed by Charles Darwin and Alfred Russel Wallace as the main mechanism behind the evolution of life (1858). Natural selection is the only adaptive process and can be defined as the differential reproduction of individuals in a population due to the presence of (dis-) advantageous genetic diversity that arises as a consequence of genetic *mutation*. Therefore, the non-adaptive process of *mutation* is the ultimate source of genetic diversity upon which natural selection can act (Lynch 2007). The other two non-adaptive processes include *recombination*, which allocates and shuffles genetic variation within chromosomes, and *genetic drift*, which ensures the random fluctuation of allele frequencies in each generation (Hartl and Clark 1989). The combined action of adaptive and non-adaptive processes in different *loci* leaves recognizable traces on the levels of genetic diversity in populations.

Population genetics encompasses the study of these different evolutionary forces in a well-defined probabilistic framework. Specifically, different approaches in population

genetics focus on understanding patterns of global (genome-wide) and local (*loci*-level) genetic diversity, both within and between biological populations.

3.2 Selection comes in different flavors

As first proposed by Darwin and Wallace (1858), natural selection relies in the principle that biological traits that allow for better survival and reproduction of carrier individuals will become more frequent in populations over time. However, the concept of natural selection has also evolved and is currently an umbrella term that defines different adaptive processes. The 'classical' definition of natural selection is therefore currently referred to as *positive selection*, in order to differentiate this from other types of adaptive processes, namely *negative* or *purifying selection*, and *balancing selection*. Contrary to positive selection, purifying selection is characterized by the elimination of genetic variants that decrease the fitness of carriers (Hartl and Clark 1989), whereas balancing selection maintains advantageous genetic diversity in populations (Andrés 2011; Key et al. 2014). The different mechanisms through which selection acts can thus result in very different patterns of local genetic diversity: on the one hand, both positive and purifying selection cause a reduction in the local effective population size (Ne), which in consequence causes a decrease in genetic diversity, respectively by allowing for the fixation of favorable alleles and linked neutral variants (selective sweep), or by purging neutral alleles as a consequence of eliminating linked deleterious variation (background selection) (Hartl and Clark 1989; Charlesworth et al. 1997). On the other hand, the action of balancing selection results in an increase of local genetic diversity (Andrés 2011; Key et al 2014). Understanding the different types of natural selection acting on the genomes of different organisms allows for a deeper understanding of the biological processes

involved in adaptation to different environments, and helps bring to light possible causes underlying the differentiation of contemporary species and populations.

3.3 Balancing selection

The neutral theory of molecular evolution states that most of the genetic variation found within (and between) populations is effectively neutral, with allele frequencies changing through time due to the action of genetic drift (Kimura 1983). Nevertheless, there are various examples of advantageous polymorphisms maintained by balancing selection in the genome (Key et al. 2014). Balancing selection acts through a variety of mechanisms including *overdominance* or heterozygote advantage, negative *frequency-dependent selection*, *temporal* or *spatial* variation in selective pressures in pammitic populations, and *pleiotropy* (Wright S. 1939; Allison 1956; Pasvol et al. 1978; Gillespie JH 1978; Gigord et al. 2001; Muelenbachs et al. 2008; Gendzekhadze et al. 2009; Hughes et al. 2013). Importantly, balanced polymorphisms are maintained in a population when the strength of selection is able to overcome the effects of genetic drift preventing the fixation of alleles (Key et al. 2014). The accumulation of neutral variation segregating close to the selected site increases local genetic diversity and results in deep local genealogies, with sequences exhibiting an older time to the most recent common ancestor (TMRCA) than expected under neutrality (Rasmussen et al. 2014). Neutral polymorphisms will segregate at frequencies close to the frequency equilibrium, which is the frequency that maximizes fitness in the population (Andrés 2011). Balancing selection thus results in patterns of increased local genetic diversity, with an excess of polymorphic over divergent sites that segregate at intermediate frequencies (Hudson and Kaplan 1988; Takahata and Nei 1990; Nordborg 1997; Barton and Etheridge 2004; Williamson et al. 2004). The

length of the genomic segment exhibiting such signatures will narrow through time and is directly dependent on the action of recombination disrupting linkage between the selected site and nearby neutral polymorphisms (Charlesworth et al. 1997; Andrés 2011; Key et al. 2014). These signatures are therefore key in uncovering loci that are potential targets of the action of balancing selection (Figure 1a).

3.4 Uncovering targets of balancing selection

The action of natural selection (*positive*, *purifying* or *balancing*) can, in principle, last for millions of years. In the case of positive selection, the selective sweep that leads variants to become fix is very difficult to detect once fixation occurs and genetic diversity reaches again equilibrium (Sabeti et al. 2002). In contrast, different methods exist that allow for uncovering genomic regions where the action of *purifying* selection imposes strong selective constraint (e.g. Ward and Kellis 2012). Finally, in the case of *balancing* selection, there are different ways to detect its signature over millions of years of evolution. For example, detecting an excess of heterozygous individuals (compared to neutral expectations and in violation of Hardy-Weinberg equilibrium) is a strong indication that a locus might be overdominant (Key et al. 2014). Combining this information with environmental factors known to impact the fitness of individuals in a population has been extremely useful in the past (Figure 1b and 1c). A classical example of heterozygous advantage in human populations is a polymorphism in the β-globin gene. Homozygous individuals for the mutant allele at this locus suffer from sickle cell anemia (HbS/HbS genotype), whereas homozygous individuals for wild-type allele (HbA/HbA genotype) are susceptible to malaria. On the contrary, heterozygous individuals are healthy and have a lower chance of becoming infected with *Plasmodium falciparum*, making them less susceptible to

malaria. Hence, in regions of the world where malaria is endemic, both alleles are maintained due to higher fitness of heterozygous individuals (Pasvol et al. 1978). Although not a trivial task, measuring genotype fitness in populations has allowed to uncover several targets of balancing selection in humans (Muelenbachs et al. 2008) and other species (Olendorf et al. 2006; Mosher et al. 2007; Fasquelle et al. 2009; Johnston et al. 2013). Additionallly, instances of balancing selection were found by correlating environmental selective pressures such as pathogen diversity with allele frequencies in human populations (Fumagalli et al. 2011) (Figure 1d).

3.5 Trans-species polymorphisms – old, but how old?

Arguably the most distinctive feature of balancing selection is the possibility to detect its action in the genome (potentially) even after millions of years after the onset of selective pressures (the same is true for *purifying* selection albeit in this case the evidence requires analyzing various genomes from different species). While it is certainly possible that recent instances of balancing selection exist in the genome, detecting these examples is very difficult, particularly because they can be confounded with ongoing selective sweeps. From this it follows that balancing selection might be actually more pervasive in the genome than is commonly conceived. Nevertheless, in cases where balancing selection is strong and constant through millions of years, selected polymorphisms present in an ancestral population may survive speciation events and segregate in present-day populations of different species, resulting in a trans-species polymorphism - trSNP (Muirhead et al. 2002; Charlesworth 2006; Andrés 2011; Key et al. 2014) (Figure 1a).

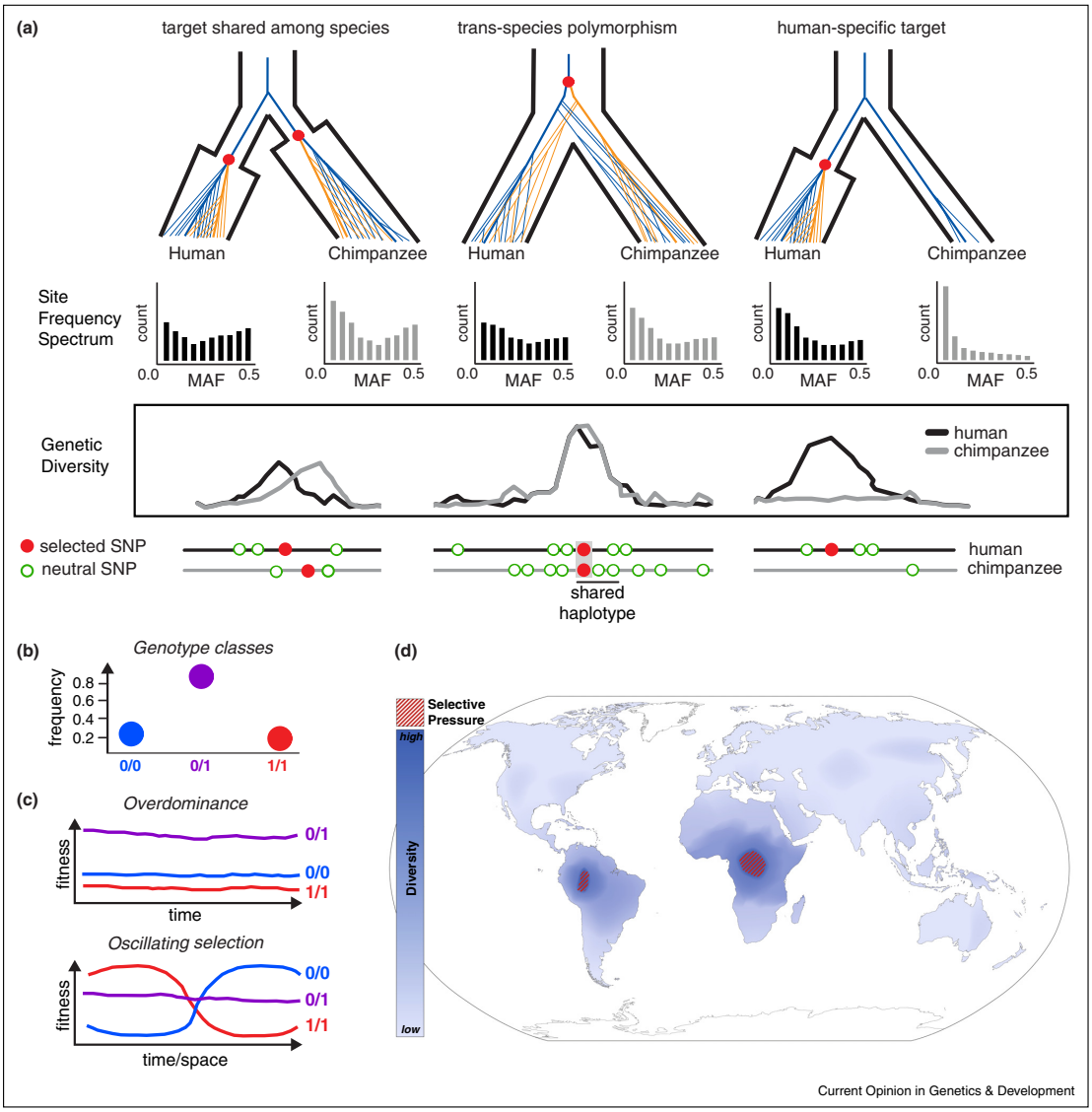


Figure 1. Strategies to identify balancing selection. (a) Using patterns of linked variation, including high genetic diversity and shifts in the folded site frequency spectrum (MAF is minor allele frequency). (b) Observing departures of Hardy-Weinberg Equilibrium (stable excess of heterozygotes). (c) Measuring fitness differences among genotype classes (e.g. overdominance and oscillating selection). (d) Detecting an unexpected correlation between genetic diversity and a given selective pressure. From *Key, Teixeira et al. (2014)*.

The presence of trSNPs arguably provides the strongest evidence for the action of long-term balancing selection in the genome, and several examples exist in a variety of organisms. The most striking example of trSNPs is the major histocompatibility locus (located on chromosome 6 – MHC) in vertebrates, where million-years old

polymorphisms were found segregating in different species of primates (Klein et al. 1993; Asthana et al. 2005; Loisel et al. 2006; Ségurél et al. 2012; Leffler et al. 2013), birds (Kikkawa et al. 2009; Sutton et al. 2013), rodents (Cutrera et al. 2007), and fish (Graser et al. 1996). The MHC region encodes proteins that have a fundamental role in immune function by presenting intracellular peptides to immune cells, which initiates immunological reactions against infected cells (Harding and Geuze 1993; Germain 1994). The trSNPs found in this region seem to play a role in the recognition and presentation of an immense variety of pathogens (Hughes and Nei 1988, 1989). Apart from the MHC region in vertebrates, others examples of trSNPs include alleles that are associated with self-incompatibility in plants (Roux et al. 2013) and heterokaryon-incompatibility in fungi (Wu et al. 1998).

3.6 trSNPs in the human lineage

The presence of trSNPs in humans has always been considered rare, mostly because the absence of unbiased, high-quality genome-wide polymorphism data in closely related species, the great apes, precluded their identification. Until recently, only a few trSNPs were described between humans and chimpanzees, most of which are located in the MHC region (Klein et al. 1993; Asthana et al. 2005). Outside the MHC region, candidate-gene approaches unveiled trSNPs between humans and chimpanzees in *TRIM5*, a gene that encodes a retroviral transcription factor (TRIM5α), and is associated with a reduced risk of HIV-1 infection (Cagliani et al. 2010), and *ZC3HAV1*, a gene associated with multiple sclerosis (Cagliani et al. 2012). More recently, it has been shown that long-term balancing selection is responsible for maintaining the trSNP defining the A and B blood groups in the *ABO* gene. This represents one of the oldest trSNPs in the genome, and segregates for tens of millions

of years in the genomes of hominoids and old-world monkeys (Ségurél et al. 2012). Finally, the first genome-wide scan aimed at uncovering trSNPs revealed the existence of six short shared haplotypes (containing at least two trSNPs) in humans and chimpanzees outside the MHC region, with the authors proposing a possible regulatory role for balancing selection due to the lack of coding trSNPs found in the study (Leffler et al. 2013).

3.7 Long-term balancing selection in humans

To date, only a few studies have attempted to unveil targets of balancing selection using a genome-wide approach (Asthana et al. 2005; Bubb et al. 2006; Andrés et al. 2009). Andrés and colleagues presented a particularly interesting one by using Sanger sequencing data on protein-coding genes in African- and European-American populations (Andrés et al. 2009). The study combined an analyses on the deviations of patterns of genetic diversity and allele frequencies from neutral expectations, and provided a catalog composed of 60 targets of long-term balancing selection in humans, among which were genes encoding keratins and membrane channels, as well as genes involved in immune response (Andrés et al. 2009). More recently, two additional studies used Complete Genomics data (Drmanac et al. 2010) to address the same question but using different strategies: the first study implemented a Discretized Sequentially Markov Coalescence (DSMC) model inferring the ancestral recombination graph in humans, and uncovered potential targets of balancing selection by looking at regions with unusually long TMRCA (some of which overlap with putatively regulatory regions in the human genome) (Rasmussen et al. 2014); the second study developed two novel likelihood ratio tests in order to identify the patterns of genetic diversity linked to a balanced polymorphism given the local

genealogies (DeGiorgio et al. 2014). Taken together, results from these studies show that, although not very common, balancing selection is likely an important force driving adaptation in humans, and targets different biological processes by acting both in coding and regulatory variation (Key et al. 2014). Particularly, targets of balancing selection seem to preferentially affect immune function (Bamshad et al. 2002; Asthana et al. 2005; Prugnolle et al. 2005; Bubb et al. 2006; Ferrer-Admetlla et al. 2008; Fumagalli et al. 2009; Andrés et al. 2009) albeit additional candidates include genes that encode proteins from the extracellular matrix, which are possibly associated with virus diversity (Andrés et al. 2009; Fumagalli et al. 2010; reviewed in Andrés 2011 and Key et al. 2014), olfactory receptors (Alonso et al. 2008), and loci associated with sperm-egg competition (Christensen et al. 2006; Hamm et al. 2007).

3.8 NCD: a novel method to detect long-term balancing selection

Recently, Bitarello and colleagues developed a new method to test the deviation of allele frequencies in a particular locus from frequencies expected under balancing selection (Bitarello et al. in preparation). The authors measure a so-called "Non-Central Deviation" (NCD) that they define as the degree to which the local site frequency spectrum SFS deviates from a pre-specified allele frequency (the *target frequency*, *tf*). After performing extensive simulations, the authors show that their method is at least as powerful as existing methods (Hudson et al. 1987; Tajima 1989; DeGiorgio et al. 2014) to detect balancing selection using realistic demographic scenarios for human populations (Bitarello et al. in preparation), and is particularly strong in uncovering old instances of balancing selection.

They propose two different implementations for this statistic: *NCD*1 and *NCD*2. The *NCD*1 statistic is based solely on the site-frequency spectrum (SFS) and uses information on allelic frequency (p_i) for each site in a locus:

$$NCD1_{tf} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - tf)^2}{n}}$$ (Equation 1 from Bitarello et al. *in prep*)

where $i = \{1,2,3,...n\}$ is the *i*-th polymorphism, and p_i is the minor allele frequency (MAF) of the *i-th* polymorphism in a locus.

The *NCD*2 statistic is an extension of *NCD*1 that uses additional information on the number of fixed differences (FDs) to an outgroup in a locus (n_{fd}), which are considered to have a MAF = 0:

$$NCD2_{tf} = \sqrt{\frac{n_{fd} \cdot (0 - tf)^2 + \sum_{i=1}^{n}(p_i - tf)^2}{n_{fd} + n}}$$ (Equation 2 from Bitarello et al. *in prep*)

3.9 Motivation

As described above, balancing selection plays an important role in the evolution of humans. However, little evidence for the action of balancing selection in shaping the genomes of our closest living relatives exists, with the exception of few targets that were identified through trSNPs between humans and great apes – particularly chimpanzees (Cagliani et al. 2010; Cagliani et al. 2012; Ségurél et al. 2012; Leffler et al. 2013). Another interesting example of long-term balancing selection in chimpanzees outside trSNPs is the gene *OAS1*, which is affects innate immune

response (Ferguson et al. 2012). Interestingly, this gene has evidence for adaptive introgression from Neandertals into modern humans (Mendez et al. 2013).

In evolutionary terms, the split between humans and other great apes is relatively recent (Figure 2), and it might be expected that most regions of the genome show similar evolutionary forces in the different species, as demonstrated by the aforementioned presence of trSNPs between humans and other apes. Nevertheless, in most cases selective pressures may shift since the split of our species. One example of this is the CC chemokine receptor 5 gene (*CCR5*), which encodes a cell-surface receptor exploited by the immunodeficiency virus type I (HIV-I) to gain entry into leucocytes. This gene shows signatures of balancing selection in humans although has undergone a selective sweep in chimpanzees (Alkhatib et al. 1996). It is therefore highly relevant to investigate whether humans and other primates tend to share evolutionary adaptations or have adopted different strategies in response to different environmental changes after diverging.

In this work we used different strategies in order to understand the extent to which targets of long-term balancing selection are shared among great ape species.

Specifically, we aimed to:

1. Identify loci with trans-species polymorphisms maintained by long-term balancing selection in the *Homo-Pan* clade that are segregating since the common ancestor of humans, chimpanzees and bonobos. For that purpose, we use high quality exome-wide data for 20 african humans, 20 central chimpanzees and 20 bonobos. Our strategy therefore allows for uncovering genes where balancing selection has

maintained an advantageous variant in these species for ~14 million years of independent evolution.

2. Determine the level of conservation of long-term balancing selection in great apes. Even in the absence of trans-species polymorphisms, balancing selection may target the same loci in different great ape lineages. Our goal is to assess the extent to which balancing selection persists in these species. To address this question, we use recently available genome-wide SNP data for 9 different great ape species, including all major clades (chimpanzees, bonobos, gorillas, and orangutans) (Prado-Martínez et al. 2013), and perform a genome analysis to identify signatures of long-term balancing selection in each species using NCD2 (Bitarello et al. in preparation). In addition, this study represents the first genome-wide scan searching for signatures of balancing selection in all great ape clades.

The combined used of the aforementioned strategies allows for the discovery of loci where balancing selection is likely millions of years old, and in some cases predates the split between different species.

4. Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees and bonobos

This manuscript is published in Molecular Biology and Evolution (MBE). Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution (SMBE) (doi:10.1093/molbev/msv007).

João C. Teixeira[1]*, Cesare de Filippo[1]*, Antje Weihmann[1], Juan R. Meneu[1], Fernando Racimo[2], Michael Dannemann[1], Birgit Nickel[1], Anne Fischer[3], Michel Halbwax[4], Claudine Andre[5], Rebeca Atencia[6], Matthias Meyer[1], Genís Parra[1], Svante Pääbo[1] and Aida M. Andrés[1]

[1]*Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany*

[2]*Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA*

[3]*International Center for Insect Physiology and Ecology, Nairobi 30772-00100, Kenya*

[4]*Clinique vétérinaire du Dr. Jacquemin, 94700 Maisons-Alfort, France*

[5]*Lola Ya Bonobo sanctuary, Kinshasa, Democratic Republic Congo*

[6]*Réserve Naturelle Sanctuaire à Chimpanzés de Tchimpounga, Jane Goodall Institute, Pointe-Noire, Republic of Congo*

*Authors contributed equally

4.1 Abstract

Balancing selection maintains advantageous genetic and phenotypic diversity in populations. When selection acts for long evolutionary periods selected polymorphisms may survive species splits and segregate in present-day populations of different species. Here, we investigate the role of long-term balancing selection in the evolution of protein-coding sequences in the *Homo-Pan* clade. We sequenced the exome of 20 humans, 20 chimpanzees and 20 bonobos and detected eight coding trans-species polymorphisms (trSNPs) that are shared among the three species and have segregated for approximately 14 million years of independent evolution. While the majority of these trSNPs were found in three genes of the MHC cluster, we also uncovered one coding trSNP (rs12088790) in the gene *LAD1*. All these trSNPs show clustering of sequences by allele rather than by species and also exhibit other signatures of long-term balancing selection, such as segregating at intermediate frequency and lying in a locus with high genetic diversity. Here we focus on the trSNP in *LAD1*, a gene that encodes for Ladinin-1, a collagenous anchoring filament protein of basement membrane that is responsible for maintaining cohesion at the dermal-epidermal junction; the gene is also an autoantigen responsible for linear IgA disease. This trSNP results in a missense change (Leucine257Proline) and, besides altering the protein sequence, is associated with changes in gene expression of *LAD1*.

4.2 Introduction

Balancing selection maintains advantageous polymorphisms in populations, preventing fixation of alleles by drift and increasing genetic diversity (Charlesworth 2006; Andrés 2011; Key et al. 2014). There are a variety of mechanisms through which balancing selection can act, including overdominance or heterozygote advantage (Allison 1956; Pasvol et al. 1978), frequency-dependent selection and rare-allele advantage (Wright 1939; Gigord et al. 2001), temporal and spatial variation in selective pressures (Gillespie 1978; Muehlenbachs et al. 2008), or pleiotropy (Gendzekhadze et al. 2009).

When balancing selection acts on a variant long enough it creates long local genealogies, with unusually old coalescence times. Selected alleles can segregate for millions of years, with neutral diversity accumulating near the selected variant(s) due to linkage (Charlesworth et al. 1997; Clark 1997; Charlesworth 2006). Selection maintains alleles close to the frequency equilibrium, the frequency that maximizes fitness in the population. This results in an enrichment of variants close to the frequency equilibrium in selected and linked variation (Hudson and Kaplan 1988; Takahata and Nei 1990; Charlesworth et al. 1997; Charlesworth 2006). Recombination restricts these signatures to short genomic segments (Wiuf et al. 2004; Charlesworth 2006; Ségurel et al. 2012; Leffler et al. 2013). If selection is strong and constant enough, the polymorphism may survive the split of different species and persist in present-day populations of more than one species, resulting in a trans-species polymorphism (trSNP) (Muirhead et al. 2002; Charlesworth 2006; Andrés 2011) (Figure 1). In species with old enough divergence time trans-species

polymorphisms are rare under neutrality and are hallmarks of balancing selection (Charlesworth et al. 1997; Clark 1997; Wiuf et al. 2004).

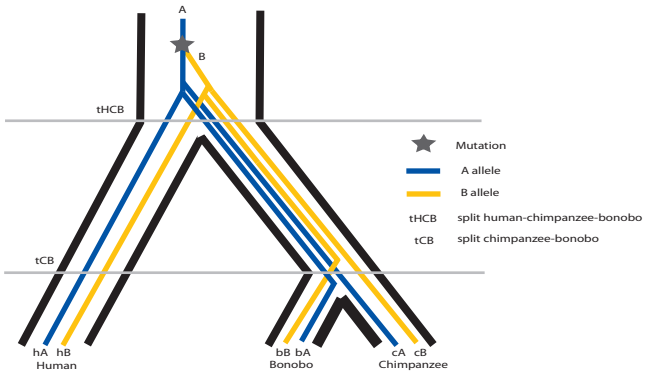


Figure 1: Schematic representation of a possible genealogy leading to a trans-species polymorphism (trSNP) in human, chimpanzee and bonobo.

The assumption that trans-species polymorphisms are very rare in humans combined with the absence of unbiased genome-wide polymorphism datasets in other great ape species resulted in few trans-species polymorphisms being described in humans: Several SNPs in the major histocompatibility locus (MHC) (Klein et al. 1993; Asthana et al. 2005), and a few non-MHC genes (e.g. TRIM5 (Cagliani et al. 2010), ZC3HAV1 (Cagliani et al. 2012), and ABO (Ségurel et al. 2012).

Recently, six well-defined short trans-species haplotypes containing at least two trSNPs shared in humans and chimpanzees have been identified (Leffler et al. 2013). Interestingly, none of these haplotypes contains coding SNPs, and the authors propose a role in the regulation of genes for the maintenance of these SNPs. Leffler et al. (2013) also identified a number of coding SNPs shared between humans and chimpanzees, but because filtering on allelic trees or CpG sites was not performed, it is unclear whether they represent trans-species polymorphisms or recurrent mutations (an important question in the identification of trSNPs, see below).

Here we analyze the exomes of 20 humans, 20 chimpanzees and 20 bonobos to identify trans-species polymorphisms present since the Homo-Pan common ancestor until the present-day population of each of the three species. By including the three species we focus only on strong balancing selection that has been maintained in the three lineages. Besides identifying coding trSNPs in several MHC genes, we also identify a novel trans-species polymorphism (rs12088790) maintained by long-term balancing selection in the gene LAD1 (ladinin-1).

4.3 Results

4.3.1 A model for neutral trSNPs in humans, chimpanzees and bonobos

As mentioned above, the presence of neutral trSNPs is unlikely when species diverged long ago. To estimate how probable a shared SNP would be in a sample of SNPs from one of the three species, we developed a model based on coalescent theory (Supplementary Information I), assuming the ancestral and the species-specific population sizes estimated in Prado-Martinez et al. (2013). Under this model, given that the lineages of bonobos and chimpanzees diverged only about 2 million years ago (Prüfer et al. 2012) and their present-day populations share polymorphisms, we expect, under neutrality, 0.85% of the SNPs in bonobos to be segregating in chimpanzees, and 4.6% of chimpanzee SNPs to also be segregating in bonobos (see Supplementary Information I). Conversely, a neutral trans-species polymorphism between *Homo* and any of the two *Pan* species is unlikely to occur by genetic drift alone: We estimate that a SNP found in a sample of humans has a probability $P_{HC} = 1.6x10^{-8}$ of being polymorphic in chimpanzees too (see also Supplementary Information I). The model also allows us to calculate the probability of observing a SNP shared by all three species (bonobo, chimpanzee and human) in a sample of human SNPs. This probability (called P_{FINAL}) is, under neutrality, approximately equal to $4.0x10^{-10}$. This is roughly 39 times lower than the probability that a SNP in humans is also polymorphic in chimpanzees (P_{HC}), illustrating the advantage of including bonobos in the comparison. Given that we observe 121,904 human SNPs, we expect about $5.0x10^{-5}$ neutral trSNPs in the three species. We note that these are actually overestimates, since coding variation is subject to purifying and background selection that produce shallower coalescent trees than neutrally evolving loci. Therefore, any

trSNP that we find is highly unlikely to have occurred under neutrality. An exploration of the behavior of the model under a range of parameters for the split times and population sizes is detailed in Supplementary Information I. We note that the parameters needed to explain the presence of neutral trSNPs in the three species are unrealistic, given our knowledge of human and great ape demographic history.

4.3.2 Idenfication of trSNPs

We sequenced the exomes of 20 Yoruba humans, 20 central chimpanzees (*Pan troglodytes troglodytes*) and 20 bonobos (*Pan paniscus*) to an average coverage of ~18X in each individual (data is very homogeneous across species in coverage and quality, see Materials and Methods). We uncovered a total of 121,904 high-quality SNPs in human, 262,960 in chimpanzee and 99,142 in bonobo. This represents a novel SNP discovery rate of ~33.54% in bonobo, ~49.29% in chimpanzee and ~2.8% in human (compared with Prado-Martinez et al. (2013) and dbSNP build 138). We focused on the 202 coding SNPs with the same two segregating alleles in the three species, the *shared SNPs* (shSNPs).

Two important confounding factors in the identification of trSNPs are genotype errors and recurrent mutations. To limit the influence of genotype errors in the form of mapping and sequencing artifacts, we conservatively removed SNPs that fall in sites that: 1) are in the upper 5% tail of the empirical distribution of coverage in at least one species; and 2) do not have high mappability (1 when using the 24mer filter (Derrien et al. 2012). We further removed SNPs that are not in Hardy-Weinberg equilibrium (HWE) with p-value (p) < 0.05 in at least one species (see Supplementary Information II). Regarding recurrent mutations, they are particularly likely in

hypermutable sites where the probability of a parallel mutation in two lineages is high. Examples of this are CpG dinucleotides (where a methylated cytosine can deaminate to a thymine and result in a C->T transition (Bird 1980; Hodgkinson and Eyre-Walker 2011)), but additional, cryptic heterogeneity in mutation rate exists (Hodgkinson et al. 2009; Hodgkinson and Eyre-Walker 2011; Johnson and Hellmann 2011).

Removing CpGs could reduce the number of recurrent mutations, but SNPs associated with CpGs represent a large fraction of SNPs in the genome (about 25% of human SNPs) and recurrent mutations can also occur at non-CpG sites (Hodgkinson and Eyre-Walker 2011). We therefore mark but consider CpG SNPs (those for which either allele results in a CG dinucleotide), and use additional lines of information to tell apart trSNPs from recurrent mutations. Specifically, SNPs that result from recurrent mutations are expected to fall in genomic regions that follow the species tree (Figure 2) because the most recent common ancestor of the genomic segment containing a human SNP falls in the human branch, predating (backwards in time) the coalescence of lineages from the different species (see previous section). On the contrary, trans-species polymorphisms create local genealogies that cluster by allele (Figure 2) because the most recent common ancestor of the genomic segment containing the trSNP predates the split of the three species (Schierup et al. 2001; Wiuf et al. 2004). Therefore, a SNP's surrounding region allows us to distinguish trSNPs from recurrent mutations.

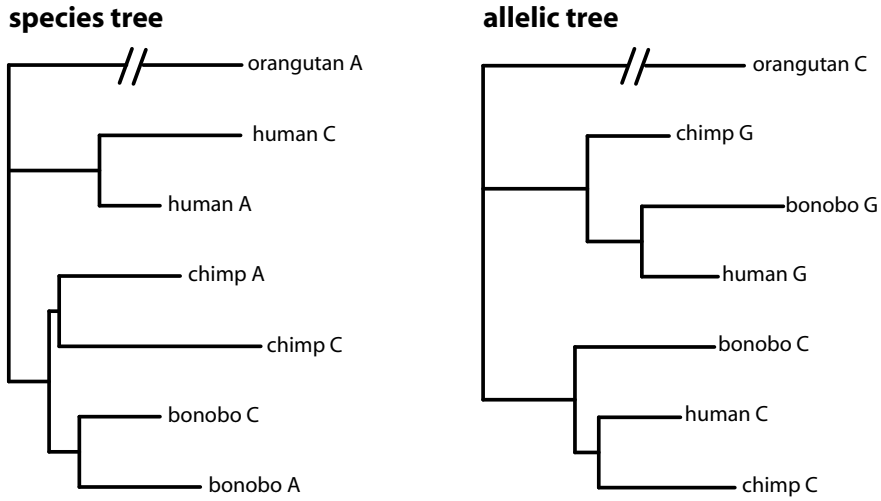


Figure 2: Examples of a species tree and an allelic tree using six haplotypes, one per species and allele. Each Neighbor-joining tree is computed on a 500 bp region around a shSNP in our dataset for the genes *TXNDC2* (species trees) and *HLA-DQA1* (allelic tree).

For each shSNP we inferred the phylogeny of its genomic region (Materials and Methods) and considered further only shSNPs that fall in genomic regions that exhibit trees that cluster by allelic type. Of the 202 original shSNPs, and after additional filtering (coverage, mappability and HWE), only 20 have a probability of an allelic tree ($P_{allelic}$) > 0.90 (Table S4); these shSNPs, all of which are present in dbSNP build 138, were considered 'candidate trSNPs'. They lie in 15 different genes, including three HLA genes. Figure 3 shows the neighbor-joining tree of one of such trSNPs, the one present in gene *LAD1*, with sequences clustering by allelic type. Only two 'candidate trSNPs' (both in *HLA-DQA1*) are not associated with CpG sites (Table S4). We also note that other HLA genes that have been described before as being targets of balancing selection in humans have shSNPs that were excluded from our analysis due to the stringent filtering criteria implemented, although no specific filters were applied in the MHC region.

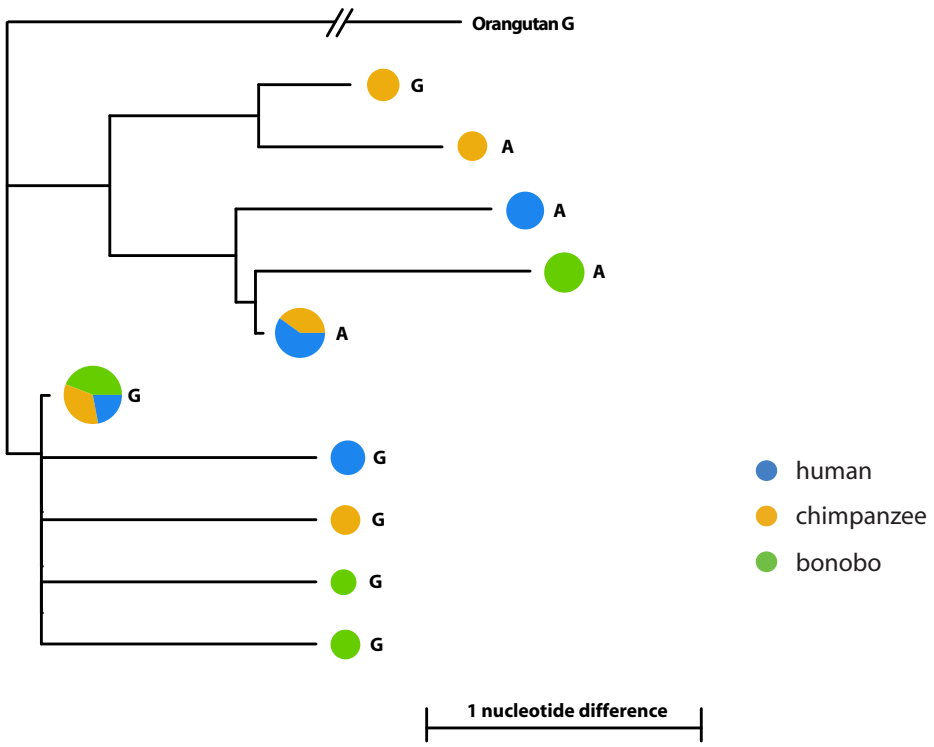


Figure 3: Neighbor-joining tree of *LAD1*. The tree was constructed using a 350 bp region as described in Methods. The size of the pie charts is proportional to the number of haplotypes (n=120), with colors representing the species. The alleles of the trSNP are shown next to the pie charts. The orangutan sequence (PonAbe2) was used as outgroup. Three chimpanzee haplotypes carrying the G allele cluster with haplotypes carrying the A allele, likely due to a recombination event (more likely to occur in chimpanzee, the species with the largest effective population size).

Because trSNPs have been previously described in HLA genes (Lawlor et al. 1988; Mayer et al. 1988; Fan et al. 1989; Klein et al. 1993; Asthana et al. 2005; Leffler et al. 2013) we focus on the remaining genes (13 'candidate trSNPs'). Our filtering criteria exclude the majority of systematic sequencing errors, so we next investigated the possibility of mapping errors due to collapsed paralogs (when paralogs are very similar in sequence, mapping errors can result in erroneous SNP calls). We BLAT (Kent 2002) the 25 bp region surrounding the 13 non-HLA candidate trSNPs to the reference genome sequences of human (hg19) and chimpanzee (PanTro4). Only four

'candidate trSNPs' (in genes *LY9*, *LAD1*, *SLCO1A2* and *OAS1*) have high mappability in all genomes (that is, a high degree of uniqueness in the genome), with the remaining nine candidate trSNPs mapping to regions that have a close paralog in at least one species (see Supplementary Information II and Table S4). Although this does not discard these positions as SNPs in the other species (or in the species with non-unique BLAT hits) we conservatively removed them from further analyses. We therefore focus on these four SNPs, to investigate additional signatures of long-term balancing selection.

4.3.3 The probability of an allelic tree

As the allelic tree provides very strong evidence for a SNP to be a trSNP, we next aim to determine how likely an allelic tree is, for each 'candidate trSNP', under recurrent mutation. To answer this question we ask how often we observe an allelic tree of the same length and minimum number of informative sites as those of the 'candidate trSNPs'. We estimated the false discovery rate (FDR, the chance of obtaining an allelic tree under recurrent mutation) for each allelic tree length by analyzing random SNPs in the genome of the three species. In short, we pair random SNPs in the human genome with a close-by SNP in chimpanzees and bonobos; these nearby, independent mutations act as pseudo-recurrent mutations where to investigate the neutral probability of an allelic tree (see Materials and Methods, SuppIementary Information III and Table S1 for details). We found that, as expected, the FDR is inversely proportional to the length of an allelic tree (Supplementary Information III); that is, the longer the genomic region, the lower the FDR of an allelic tree because the number of phylogenetically informative positions grows and so does the chance for recombination. If we condition on observing additional informative sites (besides the

'candidate trSNP') the FDR drops substantially and becomes more uniform across the different lengths (Table S1).

For our set of candidate trSNPs, and after considering the exact number of informative sites uncovered in the length of each allelic tree, only *LY9*'s allelic tree shows high FDR (36.8% for 100bp and 4 informative sites). All other candidate trSNPs fall in allelic trees that given the number of informative sites uncovered in each tree have low FDR (Table 1).

					***PtoD* (*p*)**				**MAF**		
chr:position	Gene	tree bp (FDR%)	#SNPs (H;C;B)	#FDs (H;C;B)	H	C	B	3spp	H	C	B
1:160788067*	*LY9*	100 (36.8)	(2;1;1)	(0;0;0)	0.3 (0.60)	1.1 (0.34)	0.4 (0.34)	1.7 (0.43)	0.100	0.300	0.125
1:201355761*	***LAD1***	**350 (1.5)**	**(3;3;3)**	**(0;0;0)**	**1.5 (0.02)**	**2.4 (0.06)**	**1.5 (0.02)**	**4.2 (0.03)**	**0.450**	**0.325**	**0.225**
6:31237124*	***HLA-C***	**150 (4.0)**	**(2;3;2)**	**(0;0;0)**	**25.0 (0.00)**	**22.0 (0.00)**	**20.0 (0.00)**	**38.0 (0.00)**	**0.225**	**0.225**	**0.225**
6:32609097	***HLA-DQA1***	**100 (0.0)**	**(11;7;5)**	**(0;0;0)**	**39.0 (0.00)**	**39.0 (0.00)**	**38.0 (0.00)**	**60.0 (0.00)**	**0.200**	**0.400**	**0.050**
6:32609105*		**250 (0.0)**	**(19;14;13)**	**(0;0;0)**					**0.500**	**0.350**	**0.050**
6:32609271*		**750 (0.0)**	**(25;23;24)**	**(0;0;0)**					**0.475**	**0.400**	**0.050**
6:33052736*	***HLA-DPB1***	**1000 (0.0)**	**(8;10;11)**	**(0;0;0)**	**5.3 (0.00)**	**10.3 (0.00)**	**5.8 (0.00)**	**9.8 (0.00)**	**0.300**	**0.425**	**0.325**
6:33052743		**1000 (0.0)**	**(8;10;11)**	**(0;0;0)**					**0.300**	**0.450**	**0.350**
6:33052768		**1000 (0.0)**	**(8;10;11)**	**(0;0;0)**					**0.300**	**0.475**	**0.350**
12:21453466	*SLCO1A2*	1000 (0.9)	(2;4;2)	(1;1;1)	0.6 (0.26)	1.6 (0.18)	0.9 (0.07)	2.6 (0.12)	0.025	0.350	0.050
12:113354384*	*OAS1*	250 (2.5)	(1;6;1)	(0;0;0)	0.4 (0.52)	5.3 (0.01)	0.5 (0.28)	2.3 (0.16)	0.025	0.350	0.025

* non-synonymous trSNP; H – Human; C – Chimpanzee; B – Bonobo;

Table 1: Comparison of different signatures in candidate trSNPs and genes. The estimated length of the allelic trees (and respective FDR), the number of polymorphisms (#SNPs) and fixed differences (#FDs) in the allelic tree, the polymorphism-to-divergence (*PtoD*) ratios for the whole gene and minor allele frequencies (MAF) of trSNPs are shown. For 'Human', we present the *PtoD* ratio obtained in the human-bonobo comparison, which is very similar to the human-chimpanzee comparison. The genes with trSNPs and consistent signatures of long-term balancing selection are shown in bold.

4.3.4 Excess of polymorphism linked to the trSNPs

We further investigate whether, as expected under long-term balancing selection, the 'candidate trSNPs' fall in regions that exhibit an excess of genetic diversity after taking heterogeneity in mutation rate into account. We calculated the ratio of polymorphism to divergence (*PtoD* = *p/(d+1)*, where *p* is the number of polymorphisms identified in a species and *d* the number of fixed differences identified between species – see Supplementary Information IV) in the genes containing our four non-HLA 'candidate trSNPs' (*LY9*, *LAD1*, *SLCO1A2*, *OAS1*); we also analyze the seven HLA trSNPs (*HLA-C*, *HLA-DQA1* and *HLA-DPB1*). For each gene we investigate different genomic regions, in each species: a) 'ALL' – the entire genic region; b) 'coding' – only their coding exonic sequence; c) '500bp' – the 500 bp surrounding the trSNP; and d) the 'length of allelic tree' (Table S4). First, if we focus on individual genes, *HLA* genes are in the very far tail of the empirical genomic distribution of *PtoD*, with a significant excess of polymorphism in all the comparisons performed (Table S4). For non-HLA genes, only *LAD1* shows a consistent excess of diversity in the three species, with most comparisons being significant in human and bonobo, and marginally non-significant in chimpanzee (see Tables 1, 2 and S4, and Supplementary Information IV). The weaker signal in chimpanzee is likely due to this species' larger effective population size (Prado-Martínez et al. 2013) that translates in higher genomic diversity and lower power to detect the localized increased diversity in *LAD1*. The signal is weaker for the other three genes. No excess of polymorphism is observed in *SLCO1A2*, and in *LY9* high *PtoD* values are observed only for the 'length of tree', due to the presence of a single additional SNP in humans (in such a small region). *OAS1* shows significant excess of polymorphism only in chimpanzee.

		Human	Chimpanzee	Bonobo
PtoD	ALL	1.50 (0.023)	2.40 (0.059)	1.50 (0.019)
	Coding	2.00 (0.018)	1.25 (0.332)	1.00 (0.068)
	500bp	2.00 (0.024)	1.50 (0.317)	1.50 (0.069)
	Length allelic tree	3.00 (0.028)	3.00 (0.074)	3.00 (0.024)
	3spp	4.20 (0.028)		

Table 2: *PtoD* ratios calculated in the gene *LAD1* (with the corresponding percentile in the empirical distribution in parenthesis). For 'Human', we present the *PtoD* ratio obtained in the human-bonobo comparison, which is very similar to the human-chimpanzee comparison.

We also calculate a three-species *PtoD* ('3spp') for the entire genic region by jointly considering (the union of) all polymorphisms and divergent sites across the three species. The '3spp' *PtoD* is unusually high in all HLA genes ($p \leq 0.002$) and in *LAD1* ($p = 0.028$), but not in the other three genes (Tables 1, 2 and S4). In fact, only 0.005% of genes in the genome have, in each of the three species, a p-value equal or lower than that of *LAD1*. This shows that the combined excess of diversity of *LAD1* in all three species is highly unusual. In addition, we note that *LAD1*'s signature is due to the strong enrichment in polymorphism in the region surrounding the trSNP (rs12088790): All SNPs we identified in *LAD1* are within 182bp of rs12088790.

Taken together, these results indicate that apart from the three *HLA* genes, only *LAD1* has a signature of long-term balancing selection in the three species. *OAS1* shows signatures of balancing selection in central chimpanzees, which have been previously reported (Ferguson et al. 2012), but the gene shows rather unremarkable signatures in bonobo and human (Table 1). We cannot discard the possibility that *SLCO1A2, LY9*

or *OAS1* have been under balancing selection, but conservatively we focus on *LAD1*, *HLA-C*, *HLA-DQA1* and *HLA-DPB1* as our final set of trSNPs.

The set of these four genes is, in all species, significantly more polymorphic than the empirical distribution of all genes with at least one variable site (polymorphism or substitution) in our dataset (Tables S2 and S3, and Figures S6 and S7). *LAD1* is the least polymorphic of the four genes, which is not surprising as the remaining trSNPs fall in *HLA* genes.

4.3.5 Intermediate-allele frequency of the trSNPs and linked variants

The allele frequency distribution of sites linked to a balanced polymorphism is expected to exhibit an excess of alleles at frequencies close to the frequency equilibrium. If the frequency equilibrium is high enough (e.g. 0.5) the local site frequency spectrum (SFS) will show an observable departure from the genome-wide empirical distribution. We note that the frequency equilibrium can be at any allele frequency, so while an excess of intermediate-frequency alleles is indicative of balancing selection, this is not a necessary signature.

The SFS of the four genes together shows a significant shift towards intermediate-frequency alleles, in all species (Mann-Whitney U test $p < 4\text{x}10^{-10}$; Figure 4 and Table 3). When we consider the genes individually, almost all exhibit a significant excess of intermediate-frequency alleles in all species except for *LAD1* in bonobo and human (marginally non-significant), and for *HLA-C* in bonobo (Table 3). When we combine all SNPs in each gene (the union of SNPs in all three species) and compare the resulting SFS with the combined empirical SFS (the union of all SNPs from all three

species), all genes show a significant shift towards intermediate frequencies (Mann-Whitney U test $p \leq 0.046$, Table 3), including *LAD1*.

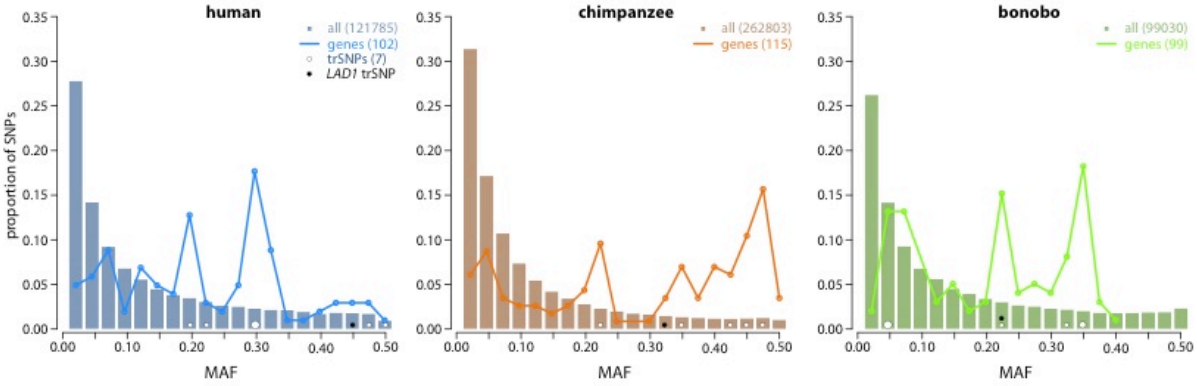


Figure 4: Folded site frequency spectra (SFS) of trSNPs and other SNPs in the genes. The x-axis represents the minor allele frequency (MAF) and the y-axis the proportion of sites in that frequency bin. The histograms show the spectrum of the entire exome ('all') for each species, excluding the four genes containing a trSNP; the lines show the combined SFS of all SNPs in the four genes containing a trSNP. The number of SNPs in each category is annotated in the legend. The trSNPs are shown as empty circles, with size proportional to the number. A black circle represents the trSNP in LAD1.

The trSNP in *LAD1*, which is a missense polymorphism, is at intermediate frequency in all three species (Table 1 and Figure 4): MAF=0.450 in human, 0.325 in chimpanzee, and 0.225 in bonobos. These frequencies are all in the upper quartile of the empirical allele frequency distributions of non-synonymous variants: In the upper 1.9% quantile for human, in the 8.6% for chimpanzee, and in the 23.8% in bonobo.

GENE	Human	Chimpanzee	Bonobo	3spp
LAD1	$5.7x10^{-2}$	$4.3\ x10^{-2}$	$7.4\ x10^{-1}$	$4.6x10^{-2}$
HLA-C	$2.5\ x10^{-2}$	$1.8\ x10^{-5}$	$5.4\ x10^{-2}$	$3.1x10^{-7}$
HLA-DQA1	$1.4\ x10^{-6}$	$1.9\ x10^{-12}$	$4.5\ x10^{-3}$	$3.4x10^{-19}$
HLA-DPB1	$4.4\ x10^{-9}$	$5.1\ x10^{-17}$	$1.2\ x10^{-11}$	$1.7x10^{-36}$
all four genes	$3.9\ x10^{-14}$	$2.0\ x10^{-30}$	$3.7\ x10^{-10}$	$1.8x10^{-54}$

Table 3: P-values (Mann-Whitney U test) for excess of intermediate-frequency alleles comparing the SFS of the genes to the genome-wide SFS.

When we investigate the 1000 Genomes dataset (Abecasis et al. 2012), which contains both coding and non-coding data for *LAD1*, we observe a significant excess of intermediate-frequency alleles in all African populations, although the signature varies across human groups with some non-African populations showing an excess of low-frequency alleles instead (Table S6). The trSNP is itself present in all these populations throughout the world: At intermediate frequency in all African populations (31% < MAF < 48%) and at lower frequency (MAF < 8%) in the non-Africans. Interestingly, when we compute F_{ST} (Weir and Cockerham 1984) values for *LAD1*'s trSNP between the African Yoruba and two non-African populations (Toscani and Han Chinese) we observe high allele frequency differences (F_{ST} = 0.238 and 0.293, respectively), which are in the 6.5% tail of the empirical F_{ST} distribution. If we condition the empirical distribution to contain only alleles observed at intermediate frequency in Yoruba (30% < MAF < 50%), the F_{ST} of rs12088790 is not in the upper tail ($0.114 < P < 0.310$ – Table S9). The polymorphism is thus shared across human populations, although its frequency shows certain differences among human groups.

4.3.6 Balancing selection in *LAD1*

LAD1 (*ladinin 1*) spans 18,704 bp and is composed of 10 exons. We obtained a total of 1,213bp of the gene by sequencing the complete exons 4, 7 and 9, as well as parts of exons 2, 3 and 5. The trSNP found in *LAD1* (chr1: 201355761, rs12088790) lies in a position that has an average mappability > 0.9 when considering 24-mers and average mappability = 1 when considering 35-mers (we note that our reads are paired end 75-mers). rs12088790 is an A/G polymorphism (reverse strand), which we validated with Sanger sequencing, and that results in a missense change located in exon 3 that results in a Leucine to Proline change. The change has a moderately conservative Grantham score (amino acid replacement score based on chemical similarity – Leucine -> Proline = 98) (Grantham 1974).

Besides altering the sequence of the protein, the trSNP is associated with expression changes in present-day humans. Specifically, when we analyzed expression data in lymphoblast cell lines from a subset of the 1000 Genomes project individuals (Lappalainen et al. 2013), we observed significantly lower expression of *LAD1* in carriers of at least one ancestral *G* allele (GG and GA genotypes) than in AA homozygotes (p = 0.02). Comparing carriers of at least one *A* allele with GG homozygotes did not show a significant difference in expression levels (p = 0.21). This shows that the derived *A* allele is associated with increased expression of *LAD1* in an at least partially recessive manner. Mapping biases are not responsible for this result as the total number of SNPs uncovered in the closest region (one additional SNP in the 150 bp region that affects read mapping) is only moderate.

4.4 Discussion

By comparing the exomes of humans, chimpanzees and bonobos, we identify polymorphisms maintained by long-term balancing selection in the Homo-Pan clade. Undoubtedly, other cases of long-term balancing selection exist, including species-specific balancing selection (Pasvol et al. 1978; Bamshad et al. 2002; Wooding et al. 2004; Wooding et al. 2005; Muehlenbachs et al. 2008; Andrés et al. 2009; Andrés et al. 2010), but here we focus on selection that is old, strong, constant and shared across lineages, and that results in trans-species polymorphisms. Even among trSNPs, we focus only on coding variants shared among the three species, and likely underestimate the number of human trSNPs. First, by focusing on coding variation we are blind to balancing selection that maintains variants outside genes, which may not be rare (Leffler et al. 2013). Second, by restricting on a SNP being present in the three species we discard cases where the variant was lost in one of the lineages, which may again not be rare. Even one of the best-established cases of trSNPs, the one present in the ABO gene from humans to old word monkeys, is not shared among the three species because it was lost in chimpanzees (Ségurel et al. 2012). This is not unexpected as it is likely that one of the species has undergone demographic or selective changes that weakened or changed selection on an old balanced polymorphism. Conversely, considering three species (e.g. adding bonobo) reduces the probability of trSNPs under neutrality; in fact, after considering the number of SNPs discovered in humans (121,904), we expect to observe no neutral trSNP (specifically, we expect 5.0×10^{-5} neutral trSNPs). Consistent with this, the majority of coding shSNPs we identified are likely the result of recurrent mutations, as they fall in genomic regions whose phylogenies agree with the expected species tree.

We identify seven trSNPs that pass our filtering criteria and that cluster by allelic tree, with an extremely low probability under recurrent mutation. The loci containing these seven SNPs show, in addition, the excess of polymorphism expected under long-term balancing selection. Six trSNPs are located in HLA genes (*HLA-DQA1*, *HLA-C* and HLA-DPB1) and one is a non-synonymous SNP in exon 3 of the gene *LAD1* (rs12088790). This variant, which has segregated for millions of years in these lineages, represents to our knowledge the only trans-species polymorphism known to segregate in present-day populations of these three species outside of the MHC. As for the remaining candidate trSNPs, the combined results of our analyses are not strong and consistent enough to provide unequivocal evidence that these are targets of long-term balancing selection (although they can be). We thus focus on *LAD1*, where the evidence is clear.

Besides containing a trSNP whose genomic region clusters by allelic type, *LAD1* exhibits high levels of genetic diversity (particularly in bonobos and humans) and it shows excess of intermediate-frequency alleles (significant in chimpanzee and marginally non-significant in humans, although highly significant in the 1000 Genomes' Africans). *LAD1* is thus an unusual gene in its consistent signatures of long-term balancing selection.

The trSNP, rs12088790, segregates at intermediate frequency in Yoruba, bonobos and chimpanzees. It is present in all 1000 Genomes human populations (Abecasis et al. 2012), although at intermediate frequency in African populations and at low frequency in non-African populations. It is not uncommon for targets of long-term balancing selection to show population differences in the allele frequency distribution

(Andrés et al. 2009), sometimes due to changes in selective pressure across human groups (de Filippo et al., in preparation). The fact that only African populations show a significant excess of intermediate frequency alleles in *LAD1,* and that F_{ST} for rs12088790 is high (although not significantly so) between African and non-African populations suggest that this might be the case for *LAD1*. We expect low F_{ST} values between populations that have identical frequency equilibria, so there might have been changes in selective pressure among human groups. Although speculative, it seems possible that some environmental pressures long shared by humans, chimpanzees and bonobos, and that still affect certain African populations, have changed in other human populations. We note nevertheless that the F_{ST} values in rs12088790 are not unusually high, so the observed population differentiation is compatible with the pure effect of genetic drift.

Although rs12088790 in *LAD1* is a good candidate to have been the target of selection (being non-synonymous and present in the three species), it is possible that it is instead maintained by linkage to an undiscovered selected trSNP, as the maintenance of several linked trSNPs is possible under long-term balancing selection (Ségurél et al. 2012). Although more detailed genomic and functional analysis on *LAD1* are needed to completely clarify this question, we explored a recently published catalog of great ape genetic polymorphism in search for additional human-chimpanzee-bonobo shSNPs (Prado-Martínez et al. 2013). Besides rs12088790 (which in that dataset also segregates in all three species), we identified one additional shSNP in the three species. This SNP (rs12035254, chr1:201349024) is intronic and downstream of exon 10, and is located about 6 kbp downstream rs12088790 (see Supplementary Information VI). The distance between the two SNPs makes it unlikely that

rs12035254 is responsible for the very localized signatures in rs12088790's genomic region. We further compared the trSNPs found in this study with a list of shSNPs between human and western chimpanzee provided by Leffler et al. (2013) but were unable to retrieve them. This is likely due to different sampling and sequencing strategies adopted in the two studies (see Supplementary Information VI). Nonetheless, Leffler et al. (2013) also reported several human-chimpanzee shSNPs in the genes *HLA-DQA1* and *HLA-DPB1*, although the specific variants are different from the ones uncovered here.

Interestingly, the two alleles of rs12088790 are associated with differences in expression levels of *LAD1*, with higher expression associated with the ancestral G allele in lymphoblastoid cell lines. This highlights the possibility that, in addition to causing an amino acid replacement, the trSNP might also have regulatory effects. Association of non-synonymous alleles with differences in gene expression is not rare (Lappalainen et al. 2013), , but we cannot discard the possibility that another, nearby variant, is responsible for the observed differences in expression.

The precise biological mechanisms leading to long-term balancing selection on *LAD1* are not known. The gene encodes a collagenous anchoring filament protein of basement membrane at the dermal-epidermal junction. The mRNA and the protein are observed in a number of tissues including the gastrointestinal system (and its accessory organs), the kidney, prostate, placenta, and one type of hematopoietic cells (Kim et al. 2014). Genes involved in cell adhesion and extracellular matrix components are enriched among candidate targets of balancing selection and among genes with intermediate-frequency alleles in pathogen-rich environments (Andrés et

al. 2009; Fumagalli et al. 2009, 2001; Key et al. 2014). This suggests that certain components of the cellular junction may benefit from the presence of functional polymorphism, perhaps as a defense against pathogens. In this context, *LAD1* may represent one of such examples.

Interestingly, genetic variation in *LAD1* is associated with linear IgA disease, an autoimmune blistering disease. The disease, which affects mostly children and elderly adults (McKee et al. 2005), is caused by the presence of circulating IgA autoantibodies that target peptides in the Ladinin-1 protein, causing an immunological reaction. This results in the disruption of the dermal-epidermal cohesion, leading to skin blistering that predominantly affects the genitalia but also the face, trunk and limbs (Ishiko et al. 1996; Marinkovich et al. 1996; Motoki et al. 1997; McKee et al. 2005). Although our understanding of the effect of the disease in different populations is biased by the fact that the disease (which is rare) has mostly been studied in Western countries, some evidence suggests that it is more common in Africa (Aboobaker et al. 1991; Denguezli et al. 1994; Monia et al. 2011). Balancing selection has been proposed to play a role in the evolution of autoimmune genes, because the inflammatory response must be precisely balanced to be effective yet moderate (Ferrer-Admetlla et al. 2008). Whether balancing selection in *LAD1* is responsible for its role in auto-immunity remains though unclear. It is possible, and perhaps more likely, that autoimmune diseases appear as consequences of diversity in proteins that is maintained by balancing selection and happen to be able to initiate pathogenic immunological reactions. Further work is necessary to discern the functional consequences and advantageous role of its balanced polymorphisms in humans and other primates.

4.5 Materials and Methods

4.5.1 DNA samples and sequencing

We performed whole-exome capture and high-coverage sequencing of 20 humans, 20 central chimpanzees (*Pan troglodytes troglodytes*) and 20 bonobos (*Pan paniscus*). Human samples belong to the well-studied Yoruba population from HapMap; bonobo and chimpanzee blood samples were collected in African sanctuaries (Lola ya bonobo sanctuary in Kinshasa, Democratic Republic Congo; and Tchimpounga sanctuary, Jane Goodall Institute, Republic of Congo, respectively) and immortalized as cell culture (Fischer et al. 2011). DNA was extracted using the Gentra Purgene Tissue Kit (Qiagen), sheared to a size range of 200 to 300 bp using the Bioruptor (Diagenode) and converted into DNA libraries for capture and sequencing (Meyer and Kircher 2010). All samples were double-indexed to prevent cross-sample contamination during the processing and sequencing of the samples (Kircher et al. 2012). Exome capture was performed using the SureSelect Human All Exon 50Mb Kit (Agilent Technologies). The kit design is based on the complete annotation of coding regions from the GENCODE project with a capture size of approximately 50 Mb. We selected all Ensembl genes (mapping uniquely to hg19) that are RefSeq genes (with good functional support) and targeted by our capture design, and selected their longest RefSeq transcript. Samples were then pooled by species and sequencing was performed on Illumina's GAIIx platform, with paired-end reads of 76bp.

4.5.2 Base calling and read mapping

Base calling was performed with Ibis (Kircher et al. 2009), and reads with more than 5 bases with a base quality score lower than 15 were discarded. Reads were aligned to

the human reference genome hg19 using BWA with default parameters. Mapping all individuals to the same reference genome prevented complications from mapping to genomes of different quality. Only reads with a mapping quality (MQ) ≥ 25 and mapping outside of known segmental duplications in the three species were considered for further analysis. Specifically, the average coverage for each individual is 18.9X in human, 17.9X in chimp and 17.9X in bonobo.

4.5.3 Genotype calling and filtering

Genotype calls were performed in the autosomes using the Genome Analysis Toolkit (GATK) *UnifiedGenotyper* (version 1.3-14) (McKenna et al. 2010). Aside from true variation, these preliminary SNP calls likely include false positives due to the presence of mismapped reads, misaligned indels and systematic errors. We used a combination of strict filters to remove such errors. SNPs were removed using the following criteria (acronyms correspond to the GATK package or fields in the VCF files):

- The depth of coverage (DP) was <8 or >100 in at least 50% of the individuals of each species. This allowed us not only to exclude positions for which the coverage depth was low, but also positions that might fall in segmental duplications not annotated in the datasets above [28-30];
- The quality score (QUAL) of the call was <50;
- There was evidence of strand bias (SB>0);

- The genotype quality (GQ) was <10 in all individuals carrying the alternative allele;
- The SNP was located within 3bp of a homopolymer with a minimum length of 5bp;
- The SNP was located within 5bp up- and down-stream of an insertion or deletion (indel) polymorphism or substitution with the human reference genome.

4.5.4 Shared SNPs as trans-species polymorphisms

Wrongly mapped reads are difficult to account for and can result in an increased false discovery of shSNPs. In order to remove undetected duplications, we further filtered shSNPs to remove sites with unusually high coverage, that are in Hardy-Weinberg disequilibrium, and that do not lie in regions of high mappability in the human genome (that is, positions that have a 24mer mappability score lower than 1, as defined by the CRG Alignability track in the UCSC browser (Derrien et al. 2012) (see Results).

4.5.5 Haplotype inference and allelic trees

We use the fastPHASE 1.4.0 software (Scheet and Stephens 2006) to infer the chromosomal phase for the alleles of each of the genes containing at least one shSNP. The inferences were performed separately for each species and for each chromosome using the default parameters of fastPHASE.

The region surrounding a trans-species polymorphism is expected to follow unusual genealogies where haplotypes cluster by allelic type rather than by species. This

occurs because the age of the balanced polymorphism predates the speciation time and, unless recombination happens, there will be no fixation of new mutations. We call these two types of phylogenies "allelic tree" and "species tree" (Figure 2). The trees were inferred in windows of different lengths (from 100 bp to 2,000 bp) centered on the shared polymorphism, as the region expected to follow the allelic tree is very short due to the long-term effects of recombination. We considered as candidate trans-species polymorphisms only shSNPs that show an allelic tree with probability ($P_{allelic}$) > 0.9 in a window of at least 100 nucleotides.

We adopted a simple resampling approach to calculate $P_{allelic}$ in the region surrounding a shSNP. We randomly created 1,000 samples of six haplotypes (one haplotype per allele and per species). For each of the 1,000 resamples we built a neighbor-joining tree using as distance matrix the number of nucleotide differences among the six haplotypes. If the three closest tips were haplotypes from the three species containing the same allele of the shSNP, it was considered an allelic tree. If the two different human haplotypes are closer to each other than to any other haplotypes, the tree was considered a species tree (the relationship between chimpanzees and bonobos was not considered because shared polymorphism can occur given their short divergence time). $P_{allelic}$ was estimated as the proportion of resampled trees that were allelic trees. Figure 2 shows an example of allelic and species trees built from six haplotypes.

We also estimated the probability to observe an allelic tree of a given length (the false discovery rates, FDRs) under recurrent mutation and based on our empirical dataset. For each observed allelic tree lengths (Table S1), we randomly chose 1,000 human SNPs and the closest SNP in chimpanzee and bonobo. We then 'paired' these SNPs

(i.e. use the allelic information of each SNP) as if they occurred in the same genomic position rather than in different positions, and calculated $P_{allelic}$ for these haplotypes (based on the alleles found in each species' SNP). Because these SNPs arose from independent mutations in each lineage, they perfectly mimic a recurrent mutation (falling at the same site) in the three species. The proportion of random samplings with $P_{allelic} > 0.9$ (i.e. the criterion used to consider the trSNP) reflects the FDR for that given length.

4.5.6 Polymorphism-to-Divergence ratios (*PtoD*)

We defined the ratio of polymorphism to divergence $PtoD = p/(d+1)$, where p is the number of polymorphisms observed in a species and d the number of fixed differences between species. For each candidate gene, we estimated significance based on the percentile of each candidate in the empirical genomic distribution of all genes.

In order to ascertain significance when comparing the set of candidate loci to the set of control loci (empirical distribution), we performed 2-tail Mann-Whitney U (MW-U) tests and used a critical value of 5%. After comparing the *PtoD* values in the two groups, we sequentially removed the top candidate gene (i.e. one gene each time) from the candidate's group and recalculated MW-U p-values maintaining the control group unaltered (see Supplementary Information IV for details).

4.5.7 Measuring expression levels in *LAD1* alleles

We analyzed lymphoblastoid cell line expression data obtained from a subset of 462 of the 1000 genomes project individuals provided by Lappalainen et al. (2013). To compute gene expression we used the aligned reads provided by Lappalainen et al.

(2013) and assigned reads with a mapping quality (MQ) ≥ 30 to protein coding genes by overlapping the read coordinates with gene coordinates (ENSEMBL version 69). Reads overlapping a gene are summed up and used as the estimate for gene expression.

We grouped the individuals by their genotype at position chr1:201355761 (rs12088790, the non-synonymous trSNP in *LAD1*). We sought to test for allele-specific expression for *LAD1* between individuals carrying the two different trSNP alleles by testing for differential expression between (i) the groups of individuals with genotype AA vs. GG/GA and (ii) the groups of individuals with genotype GG vs. GA/AA. We computed differential expression for *LAD1* for (i) and (ii) using the DESeq package (Anders and Hubers 2010). Expression values in both groups are modeled by a fit of a negative binomial distribution. DESeq tests then for differences between the distributions of the two groups.

4.6 Supplementary Information

I – Trans-species Polymorphism due to Neutral Identity by Descent

In a coalescent genealogy, the probability of a neutral polymorphism shared between humans and chimpanzees due to identity by descent is very low (Leffler et al. 2013). This probability depends on at least two human lineages and two chimpanzee lineages not coalescing before the human-chimpanzee split time, and on a particular order of coalescence events in the ancestral population along with the occurrence of a mutation on the correct part of the genealogy (Wiuf et al. 2004; Ségurél et al. 2012). We targeted SNPs evolving under long-term balancing selection by focusing on trans-species polymorphisms shared between humans, chimpanzees and bonobos, so it is necessary to calculate the probability that such polymorphisms are neutral.

For this, we begin by studying the properties of a trans-species polymorphism in a coalescent genealogy of 2 lineages per species. First, looking backwards in time, for a trSNP to occur, a general requirement is that none of the pairs of lineages of each species coalesce during their species-specific history. Assuming this is the case, there are three different types of scenarios that could result in a polymorphism shared between the three species: a) none of the two chimpanzee and two bonobo lineages coalesce from the time of the chimpanzee-bonobo split to the time of the human-chimpanzee-bonobo split; b) a single chimpanzee lineage coalesces with a single bonobo lineage during this time; and c) a single chimpanzee lineage coalesces with a single bonobo lineage, and a different chimpanzee lineage coalesces with a different bonobo lineage during this time.

These different scenarios are shown in figure S1. Moreover, it is then necessary for a mutation to occur in the correct lineage in the ancestral population, such that it leads to a pattern consistent with a trans-species polymorphism. The probability of occurrence of such a mutation varies according to the number of lineages reaching the human-chimpanzee-bonobo ancestral population. In other words, different tree topologies might have different probabilities of producing a neutral trans-species polymorphism.

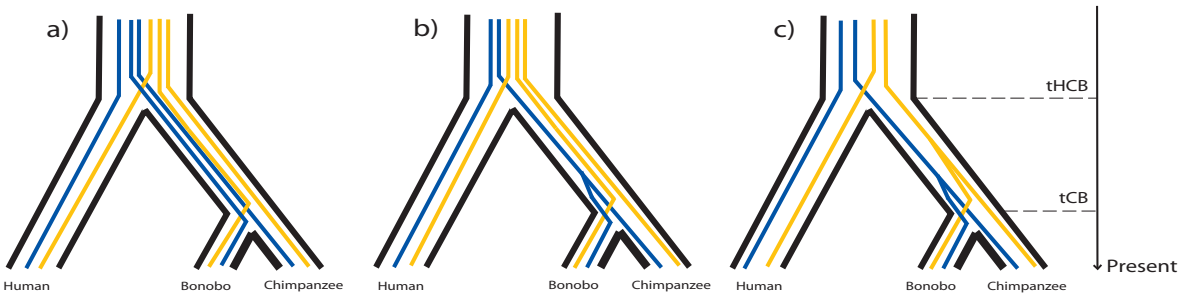


Figure S1 – Genealogical scenarios allowing for a trans-species polymorphism shared between humans, chimpanzees and bonobos. In all three examples, the mutation must arise in the ancestral population before the human-chimpanzee-bonobo split time. The blue and yellow coloring of different lineages is arbitrary, and does not denote derived or ancestral states. tHCB and tCB represent the split times of human-chimpanzee-bonobo and chimpanzee-bonobo, respectively.

Below, we estimate the probability that a human polymorphism is also a trans-species polymorphism in chimpanzees and bonobos, under neutrality. We assume that population sizes stay constant within each species (but not necessarily across species), that there is no population structure within species or migration among species, that there is no recurrent mutation, and that the number of sampled chromosomes is small relative to the whole population for each species. The probability we obtain will not depend on the value of the mutation rate per site (so

long as the mutation rate is constant along the genealogy, which we assume for simplicity), as it will be a ratio of two terms which both contain the mutation rate, and so the rate will cancel out. We first define the following terms:

Term	**Definition**
N_A	Human-chimpanzee-bonobo ancestral population size
N_H	Human population size since the population split with chimpanzees and bonobos
N_C	Chimpanzee population size since the chimpanzee-bonobo population split time
N_B	Bonobo population size since the chimpanzee-bonobo population split time
N_{CB}	Population size of chimpanzees and bonobos after the split with humans but before the split between each other
t_{HCB}	Population split time (in generations) of humans and chimpanzees+bonobos
t_{CB}	Population split time (in generations) of chimpanzees and bonobos
t_X	$(t_{HCB} - t_{CB})/(2*N_{CB})$
$P_{Htsp}(x,x+\Delta x)$	Probability of finding a site where 2 human chromosomes are different, 2 bonobo chromosomes are different and 2 chimpanzee chromosomes are different in a region of length Δx
$P_{Hhum}(x,x+\Delta x)$	Probability of finding a site where 2 human chromosomes are different in a region of length Δx
$P_{Ptsp}(x,x+\Delta x)$	Probability of finding a site where humans are polymorphic, bonobos are polymorphic and chimpanzees are polymorphic in a region of length Δx
$P_{Phum}(x,x+\Delta x)$	Probability of finding a site where humans are polymorphic in a region of length Δx
P_{TSPHET}	Probability that 2 bonobo chromosomes are different and 2 chimpanzee chromosomes are different at a site, given that 2

	human chromosomes are different at that site
P_{FINAL}	Probability that a site is polymorphic in both bonobos and chimpanzees, given that it is polymorphic in humans
u	Mutation rate per site per generation
g(n,j,t)	Ancestral process of the coalescent: probability of there being j lineages at time t in the past, given that there were n lineages at time 0, measuring time in coalescent units (Tavaré 1984)

We define ETA[k] to be the expectation for the inter-coalescence time (in generations) while there are k lineages in the human-chimpanzee-bonobo ancestral population:

$$\text{ETA}[k] = \frac{2N_A}{\binom{k}{2}}$$

We also define ETH[2, N_H, N_A, t_{HCB}] to be the expectation for the time until coalescence (in generations) of 2 lineages sampled in the human population in the present (Grifiths and Tavaré 1984):

$$\text{ETH}[2, N_H, N_A, t_{HCB}] = 2N_H \int_0^\infty e^{-\int_0^v f(z,N_H,N_A,t_{HCB})dz} dv$$

where f(z, N_H, N_A, t_{HCB}) is a piecewise constant function defined as 1 when z <= $t_{HCB}/(2N_H)$ and N_H/N_A when z > $t_{HCB}/(2N_H)$.

We begin by obtaining the probability that a site that is heterozygous in 2 human chromosomes is also heterozygous in 2 bonobo chromosomes and in 2 chimpanzee chromosomes:

$$P_{TSPHET} = \lim_{\Delta x \to 0} \frac{P_{Htsp}(x, x+\Delta x)}{P_{Hhum}(x, x+\Delta x)} = \frac{(e^{-t_{HCB}/(2N_H)})(e^{-t_{CB}/(2N_C)})(e^{-t_{CB}/(2N_B)})PX}{u*2*ETH[2, N_H, N_{A,} t_{HCB}]}$$

where $PX = [g(4,4,t_X)*PA + g(4,3,t_X)*(2/3)*PB + g(4,2,t_X)*(2/7)*PC]$

$$PC = \frac{2u*PO}{9}$$

$$PB = \frac{u}{10}\left[ETA[4] + PO\right]$$

$$PA = \frac{4}{5}PB$$

and

$$PO = (3*ETA[3] + 2*ETA[2])$$

Here, PA, PB and PC correspond to the probabilities of a mutation occurring in the correct lineage in the human-chimpanzee-bonobo ancestral population, given

scenarios a), b) and c), respectively. The ancestral process functions g(n,j,t) in each term of PX allow us to calculate the probability of each scenario.

Following Leffler et al. (2013), we can approximate the probability that a site that is polymorphic in a human sample is also polymorphic in a sample of bonobos and a sample of chimpanzees in the following way:

$$P_{FINAL} = \lim_{\Delta x \to 0} \frac{P_{Ptsp}(x, x+\Delta x)}{P_{Phum}(x, x+\Delta x)} \approx \frac{(e^{-(t_{HCB}-2N_H)/(2N_H)})(e^{-(t_{CB}-2N_C)/(2N_C)})(e^{-(t_{CB}-2N_B)/(2N_B)})PX}{u*2*ETH'[N_H, N_A, t_{HCB}]}$$

where $ETH'[N_H, N_A, t_{HCB}] = 2N_H \int_0^\infty e^{-(\int_0^v f(z,N_H,N_A,t_{HCB})dz)+1} dv$

We fixed the relevant population size and split time parameters at the values estimated in (Prado-Martínez et al. 2013) : N_A = 55,000, N_H = 8,000, N_C = 30,000, N_B = 5,000, N_{CB} = 30,000, t_{HCB} = 250,000, t_{CB} = 40,000.

Using these values, we obtain that P_{FINAL} is equal to 4.05×10^{-10}. Although we have not assessed the effect of within-species population size variation on this value, the addition of bottlenecks would only make the probability of observing trSNPs smaller, not larger, as it would increase the rate of within-species coalescence.

We can compare the obtained probability to the probability of seeing a polymorphism in a sample of chimpanzees, given that the site is polymorphic in a sample of humans, as in Leffler et al. (2013). Let us denote this probability as P_{HC}:

$$P_{HC} \approx \frac{(e^{-(t_{HCB}-2N_H)/(2N_H)})(e^{-(t_{CB}-2N_C)/(2N_C)})(e^{-(t_{HCB}-t_{CB})/(2N_{CB})})PC}{u*2*ETH'[N_H, N_A, t_{HCB}]}$$

Using the same fixed parameters as above, we obtain that this probability equals 1.58×10^{-8}, which is 39 times larger than P_{FINAL}.

Additionally, using an analogous calculation to P_{HC} and assuming $N_A = N_{CB} = 30{,}000$, we obtained the probability that a site is polymorphic in bonobos given that it is polymorphic in chimpanzees (= 0.0085) as well as the probability that a site is polymorphic in chimpanzees given that it is polymorphic in bonobos (= 0.046). The probability is higher in the latter case because $N_B < N_C$. This implies that the denominator in the first case is larger than in the second case, while the numerator stays the same.

We can also vary some of these parameters and observe the behavior of P_{FINAL} under different input values. For example, we plotted P_{FINAL} in Figure S2 as a function of the human population size (ranging from 0 to 20,000) and the chimpanzee-bonobo population size during their shared history (ranging from 0 to 200,000). As expected, as population sizes increase (making recent coalescences less likely) this probability also increases. Interestingly, the $\log_{10}$ of this probability drops sharply when either of the two population sizes are small (~1,000), because coalescent events tend to happen very early in populations of those sizes, and so trans-species polymorphisms become extremely unlikely.

In Figure S3, we show P_{FINAL} as a function of the human-chimpanzee-bonobo split time (t_{HCB}, ranging from 100,000 to 300,000 generations) and the chimpanzee-bonobo split time (t_{CB}, ranging from 0 to 100,000 generations). Again, as expected, this probability decreases as a function of the split times.

According to our approximation, the probability of a segregating site in humans being a trans-species polymorphism with chimpanzee and bonobo is 4.05×10^{-10}. Given that we find 121,904 SNPs in humans, and assuming independence between SNPs, we can use a binomial distribution with this probability to model the number of trSNPs we should observe. We expect approximately 0.00005 (basically none) shSNPs with chimpanzee and bonobo under neutrality. The probability that there is at least one neutral trSNP in our sample is also approximately equal to 0.00005.

We can also change the population sizes in the model to see how much they would need to change to reach values in the same order of magnitude as the number of candidates in our data. For example, doubling the population sizes of all three terminal branches as well as the bonobo-chimpanzee ancestral population (N_H = 16,000, N_C = 60,000, N_B = 10,000, N_{CB} = 60,000) would result in an expected number of trSNPs equal to 3.56 (P[at least 1 trSNP] = 0.97), keeping all other parameters equal. Similarly, increasing the ancestral population size by 5 orders of magnitude ($N_A = 5.5 * 10^9$) but keeping all other parameters equal would result in an expectation of 4.54 trSNPs (P[at least 1 trSNP] = 0.99). All these parameter choices are highly unrealistic. Hence, these results suggest that any trSNPs we observe are unlikely to arise by neutrality, and other forces, like long-term balancing selection, must be responsible for their maintenance.

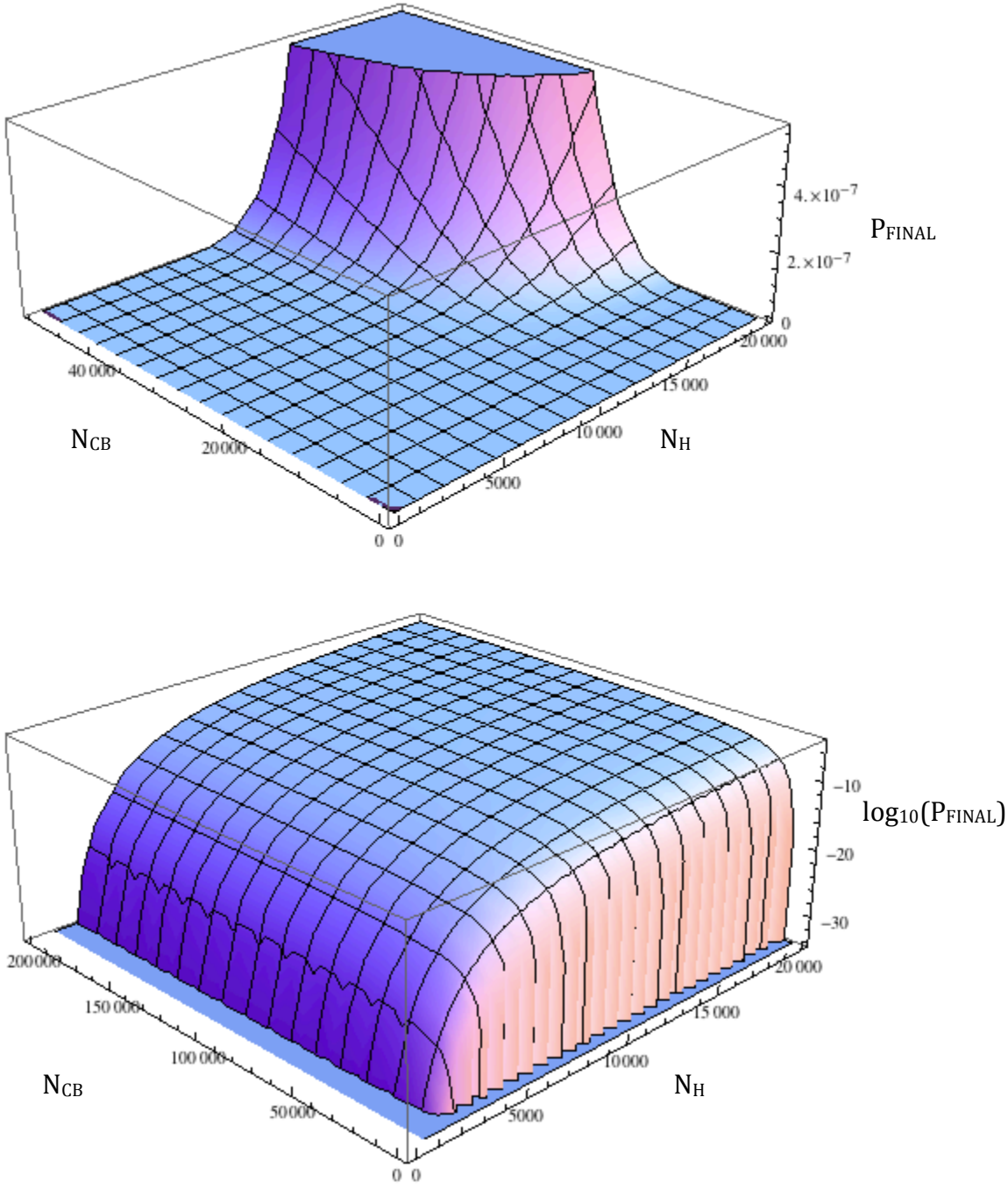


Figure S2. Top panel: P_{FINAL} plotted as a function of the human population size (ranging from 0 to 20,000) and the chimpanzee-bonobo population size during their shared history (ranging from 0 to 200,000). Bottom panel: $\log_{10}(P_{FINAL})$ plotted as a function of the same parameters. All other parameters are held fixed at the values estimated in Prado-Martinez et al. (2013).

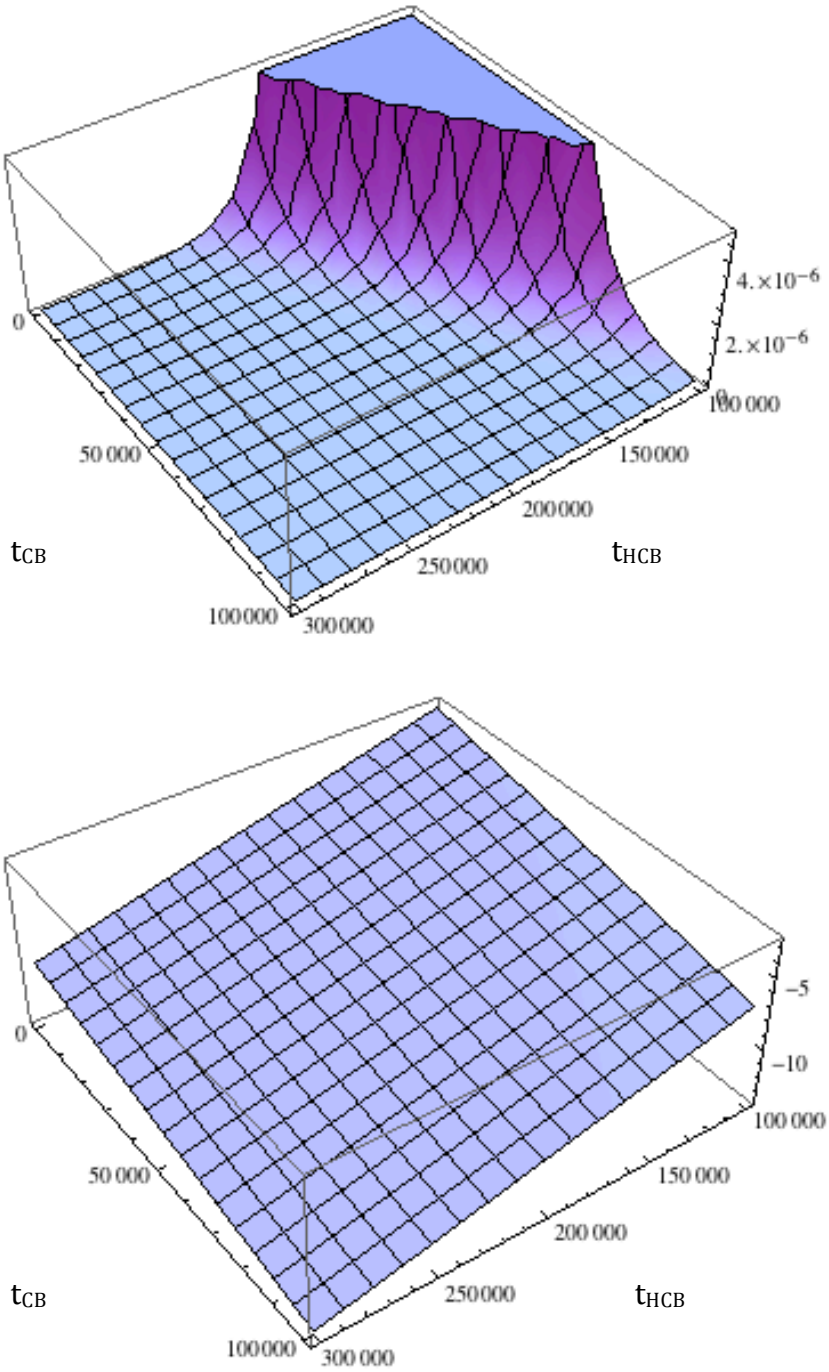


Figure S3. Top panel: P_{FINAL} plotted as a function of the human-chimpanzee-bonobo split time (t_{HCB}, ranging from 100,000 to 300,000 generations) and the chimpanzee-bonobo split time (t_{CB}, ranging from 0 to 100,000 generations). Bottom panel: $\log_{10}(P_{FINAL})$ plotted as a function of the same parameters. All other parameters are held fixed at the values estimated in Prado-Martinez et al. (2013).

We also simulated 10 sets of 10,000 genealogies for different demographic scenarios in ms (Hudson 2002) to verify our analytical expression for P_{TSPHET} was correct (Figure S4). For each set, we obtained the average P_{TSPHET} across genealogies. In the different scenarios, we used shorter population split times than in the human-chimpanzee-bonobo scenario due to the computational cost of obtaining a branch where a trans-species polymorphism can appear when population split times are far in the past. The simulated and analytical values differ most when the true value is small (e.g. Scenario C), because in those cases most of the sampled simulated genealogies contain no branches where a trans-species polymorphism is possible and so sparse sampling of the correct genealogies increases the error in the simulation estimates. Details of models A-F can be found in the caption of figure S4.

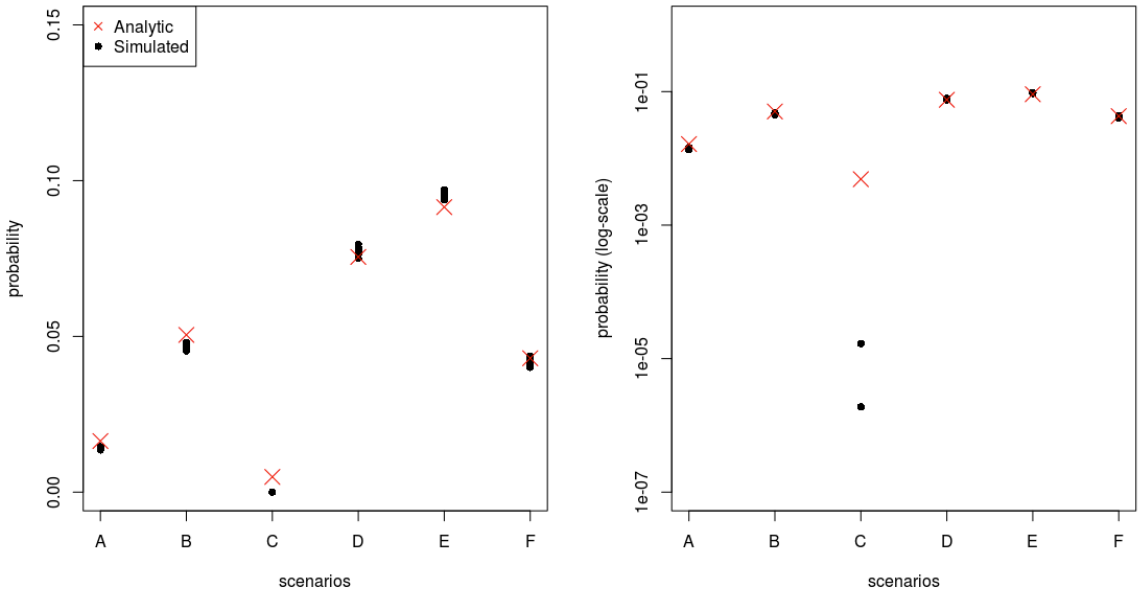


Figure S4. Analytic and simulated values for P_{TSPHET} under different demographic scenarios. The simulated values were obtained from the average of a set of 10,000 simulated genealogies, and we plotted 10 sets per scenario. The right panel shows the same values as the left panel but with a log-scaled probability on the y-axis. The parameters used for each simulated scenario were as follows: A)

$N_H = N_C = N_B = N_{CB} = N_A = 10{,}000$; $t_{HCB} = 20{,}000$; $t_{CB} = 5{,}000$. B) $N_H = N_C = N_B = N_{CB} = N_A = 10{,}000$; $t_{HCB} = 10{,}000$; $t_{CB} = 1{,}000$. C) $N_H = N_C = N_B = N_{CB} = N_A = 10{,}000$; $t_{HCB} = 30{,}000$; $t_{CB} = 10{,}000$. D) $N_H = 10{,}000$; $N_C = N_B = N_{CB} = N_A = 50{,}000$; $t_{HCB} = 20{,}000$; $t_{CB} = 5{,}000$. E) $N_H = 10{,}000$; $N_C = N_B = 50{,}000$; $N_{CB} = N_A = 100{,}000$; $t_{HCB} = 20{,}000$; $t_{CB} = 5{,}000$. F) $N_H = 8{,}000$; $N_{CB} = N_C = 30{,}000$; $N_B = 5{,}000$; $N_A = 55{,}000$; $t_{HCB} = 20{,}000$; $t_{CB} = 5{,}000$. In all but two of the sets of Scenario C, all trees simulated under this scenario contained no branches where a trans-species polymorphism is possible and so sparse sampling of simulations leads to underestimation of the true value for P_{TSPHET}. The other 8 sets therefore had an average simulated $P_{TSPHET} = 0$. The right-hand plot only shows values of average P_{TSPHET} for the two sets in Scenario C where at least one tree contained a trans-species polymorphism (average simulated $P_{TSPHET} > 0$).

II – Identification, filtering and validation of shSNPs

We performed genotype calling with GATK (McKenna et al. 2010) and proceeded to filter putative false positive variants (see Materials and Methods). We initially identified shared SNPs (shSNPs) as orthologous SNPs that showed the same two alleles in all three species. This set did not include orthologous polymorphic positions for which at least two species showed different alternative alleles, which we instead define as coincident SNPs (cSNPs) as the position is polymorphic across these species but the alleles are different.

We uncovered a total of 202 coding shSNPs in the three species. Because shSNPs might be enriched for sequencing errors, we adopted additional filtering criteria only on shSNPs to exclude such errors. Specifically, due to potential problems arising from incorrectly mapped reads, we excluded shSNPs: 1) that fall in regions with unusually high coverage (5% upper tail of the coverage distribution of all SNPs); and 2) that are not located in high mappability regions defined by 24mer mappability track (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability). In addition, we excluded shSNPs in Hardy-Weinberg disequilibrium ($p < 0.05$). Although these hard cutoffs potentially resulted in the removal of some true positives, they largely remove false SNP calls (see section on Sanger sequencing validation below). We also flagged the shSNPs that fall in a CpG site, as these are likely to be enriched in recurrent mutations. We consider as CpG sites all SNPs for which either allele results in a CG dinucleotide.

Finally, shSNPs may also be the result of recurrent mutation in the different lineages. Because a true trans-species SNP (trSNP) must fall in a region where sequences cluster by allele rather than by species, we only considered as candidate

trans-species SNPs (trSNPs) those shSNPs that show a phylogeny where haplotypes cluster in an allelic tree and not in a species tree (see Materials and Methods, and Results).

We obtained of 20 candidate trSNPs in all three species (Figure S5). Because of the divergence time between human and the two Pan species we do not expect to observe any trSNP under neutrality (see Supplementary Information I), and these are strong candidate targets of long-standing balancing selection. Of these 20 candidate trSNPs, 10 (50%) result in a non-synonymous change and alter protein sequence. 7 candidate trSNPs (35%) are located in three *HLA* genes that belong to the MHC region on chromosome 6, which is the best-established example of balancing selection in vertebrates (Klein et al. 1993; Graser et al. 1996; Asthana et al. 2005; Loisel et al. 2006; Cutrera and Lacey 2007; Kikkawa et al. 2009, Leffler et al. 2013; Sutton et al. 2013) (Table S4). Because HLAs is a known target of selection we focus below on the non-HLA candidate trSNPs.

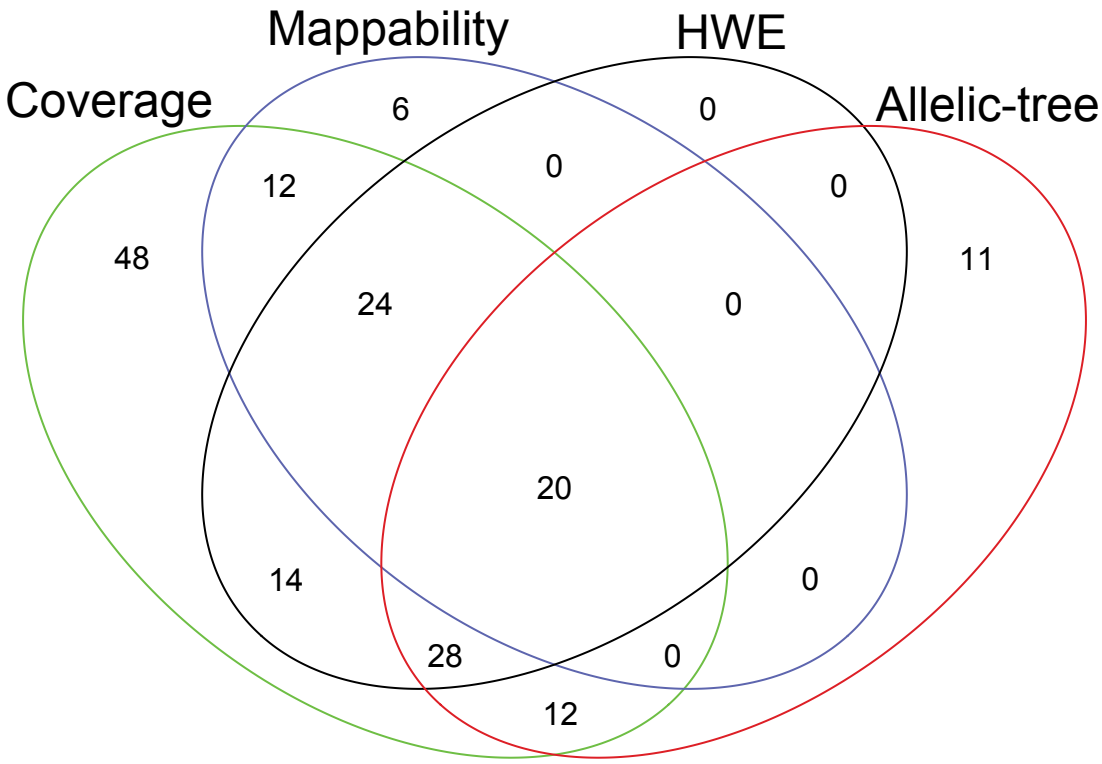


Figure S5: Venn diagram showing the number of shSNPs that passed the different filtering criteria.

a) Sanger sequencing validation

We produced Sanger resequencing data for regions surrounding regions of interest in all three species. A total of 18 bonobos, 19 chimpanzees and 18 humans were used in this analysis. The primers were designed specifically for each species by taking into account the substitutions identified in our dataset. Primer pairs were designed using Primer3 (Rozen and Skaletsky 2000), ensuring a single amplification product for the majority of the fragments (amplicon sizes vary from 504 bp to 642 bp). Additional sets of primers and different primer combinations were used in cases where a PCR reaction failed or multiple bands prevented effective sequencing. PCR reactions were performed using Herclase II Fusion

(Agilent Technologies), and following manufacturer's recommendations. After amplification, PCR products were purified using SPRI beads. Sequencing reactions were carried using the BigDye terminator v1.1 Cycle Sequencing chemistry (Applied Biosystems), and purified by ethanol/sodium acetate precipitation. Sanger sequencing was performed using an ABI 3730 sequencer (Applied Biosystems). All sequences were analyzed using the Sequencing Analysis software provided with the instrument (Applied Biosystems). We were able to validate the trSNP in *LAD1* (chr1: 201355761) in the three species.

We also attempted to validate some additional shSNPs that did not pass the Hardy-Weinberg equilibrium and mappability filters. Of these, human showed the highest percentage of validated SNPs (36.21%) and the lowest percentage of defined false positives (34.48%), with 29.31% of SNPs that we could not ascertain with Sanger, and remained ambiguous. 10.35% of SNPs were validated in bonobo, the same as in chimpanzee; moreover, 60.34% of the SNPs could not be validated and 29.31% could not be ascertained in bonobo, and in chimpanzee 46.55% were not validated 43.10% were not ascertained. Sanger validation was hampered by two main problems: first, the difficulty to obtain clean bands of expected size by PCR in some of the species; second, the presence of short non-annotated segmental duplications in some species. Specifically, in about 50% of the SNPs that failed validation we observed repeated regions of variable length (20-50 bp) around the SNP, which hampers Sanger validation. This is not surprising as these SNPs did not pass all our quality and mappability filters above, and highlights the relevance of very strict mappability and data quality filtering criteria when analyzing shSNPs.

b) BLAT analysis

We performed a BLAT analysis as a final step in order to ensure the candidate trSNPs were real using UCSC's 'BLAT Search Genome' tool. We used the -/+25 bp surrounding each trSNPs as query, and performed a BLAT of each sequence against the human genome (hg19), first using the reference allele and then using the alternative allele. After this, we repeated the analysis using the chimpanzee genome (PanTro4). This was done for all 13 non-HLA candidate trSNPs. We found that the region surrounding nine of these trSNPs is duplicated in at least one species, which increases the probability of these positions being false SNPs due to small, non-annotated duplications and mapping errors (Table S4). Conservatively, we excluded them from further analyses and focused on the remaining 11 trSNPs (7 of which are present in HLA genes). These trSNPs lie in the genes *LY9*, *LAD1*, *SLCO1A2*, *OAS1*, *HLA-C*, *HLA-DQA1* and *HLA-DPB1*.

III – False discovery rate of allelic trees

A key feature of trans-species polymorphisms is that they are expected to lie in genomic regions that form haplotypes clustering by allele rather than by species, because the two haplotypes defined by the trSNP predate species splits, unless recombination has disrupted them. Because of the long-term effects of recombination, we expect this signature to be restricted to a very short genomic region around the trSNP (Charlesworth 2006). shSNPs due to recurrent mutations, on the other hand, are expected to lie in genomic regions whose phylogeny reflects the history of the species. We take advantage of this very specific signature to identify trSNPs by focusing exclusively on shSNPs that exhibit haplotypes clustering by allele.

Because the presence of a shSNP in the absence of other informative sites can result in an allelic tree, we aim to assess how likely it is to obtain an allelic tree of a given length in any region of the genome containing a shSNP. Because shSNPs are unusual (and enriched for technical artifacts), we focus on regions of the genome that contain a human SNP and a nearby SNP in chimpanzee and bonobo. Because the mutations that lead to the human SNP and to the chimpanzee/bonobo SNP are independent, the process mimics perfectly a recurrent mutation (but affecting a different site in each lineage, rather than in the same site). We 'paired' the human and chimpanzee/bonobo SNPs and built neighbor-joining trees. We then proceeded the same way as when analyzing trSNPs (Materials and Methods). This allowed us to estimate the probability of an allelic tree under recurrent mutation (the false discovery rate, FDRs) for genomic regions of different lengths and different number of informative sites. The results are shown in Table S1.

	FDR (%)	
length allelic tree (bp)	**f1**	**f2**
1000	4.63	0.84
950	4.5	1.16
750	5.5	1.48
350	**12.61**	**1.84**
300	15.58	1.34
250	16.58	1.6
150	26.09	2.29
100	36.83	4.33

Table S1: False discovery rates (FDRs) obtained for allelic trees of the same length as the ones observed in trSNPs. Bold green shows the FDR for 350 bp, which is the length of the allelic tree in *LAD1*. Different FDRs correspond to different filters applied in the analysis: f1 – no filters; f2 – having at least another informative site (i.e. another SNP in any species and/or a fixed difference in at least one comparison HB, HC, BC).

First, using no filters we observe a clear negative correlation between FDRs and length of the allelic tree (*f1* in Table S1). This is expected given that shorter trees are likely to have fewer informative sites than longer trees, and thus to have allelic trees just as a result of the 'shSNP'.

When we condition on the presence of another informative site (SNP or fixed difference) in the region, the FDRs are more uniform across lengths (nevertheless with shorter regions showing higher FDRs – *f2* in Table S1).

If we focus on the FDRs obtained for windows with 350bp (the length observed in *LAD1*), we obtain 12.61% with no filters, and 1.84% if we condition on the windows having one additional informative site. The 350 bp region has, in *LAD1*, three additional informative sites (all SNPs). Taken together, these results strengthen the evidence that the 350 bp-long allelic tree defined by the trSNP in *LAD1* can hardly be explained by random chance.

IV – Ratio of polymorphism to divergence in candidate genes

The patterns of diversity in a region surrounding a balanced polymorphism can be used to determine whether a given locus evolved under selection. In the particular case of long-standing balancing selection, the coalescent times of selected loci will be older than those of neutrally evolving ones, which, considering a constant mutation rate, results in an excess of polymorphism and deficiency of divergence linked to the selected variant (Charlesworth 2006). We calculated the polymorphism-to-divergence ratio $PtoD = p/(d+1)$, where p is the number of polymorphisms found in a species and d the number of fixed differences between species. This statistic allowed us to infer whether the set of candidate genes was significantly more polymorphic when compared to control genes (empirical genomic distribution) and, at the same time, control for heterogeneity in the mutation rates (since both SNPs and substitutions – are included).

PtoD ratios were calculated for all genes considered as informative (i.e. all the genes that had at least one SNP or one substitution in our dataset after data quality filtering). This served as the empirical genomic distribution of *PtoD* and allowed us to quantify how diverse is the set of candidate genes in our analysis, when

compared with the empirical distribution. We calculated *PtoD* ratios in 5 different ways:

- 'ALL' (the entire set of SNPs found in the gene),
- 'coding' (only coding SNPs found in the gene),
- '500bp' (all SNPs found in the +/- 250 bp window surrounding a trSNP). In this case, and for genes that have more than one shSNP, the *PtoD* value represents the average of *PtoD* values obtained for the 500bp windows around each of the trSNPs. So assuming one single SNP is under selection, this is likely an underestimate of the diversity in the 500 bp region around the trSNP maintained by long-term balancing selection,
- 'length of allelic tree' (the surrounding regions around a trSNP that cluster by allele), and
- '3spp' (the union of all informative sites in the three species together, for the whole gene).

PtoD values were calculated separately for each of the three species. To calculate polymorphism (p) we considered the number of SNPs found in each species. To calculate divergence (d, the number of fixed differences) we proceeded as follows: i) for bonobo and chimpanzee, we used the number of substitutions relative to human; ii) for human, we performed two separate comparisons using the number of substitutions relative to bonobo and to chimpanzee, separately. The results are shown in Figures S6 and S7, and in Tables S2, S3 and S4.

As for individual genes, the pattern is dominated by HLA candidates, which is not surprising as these represent some of the most diverse genes in the genome.

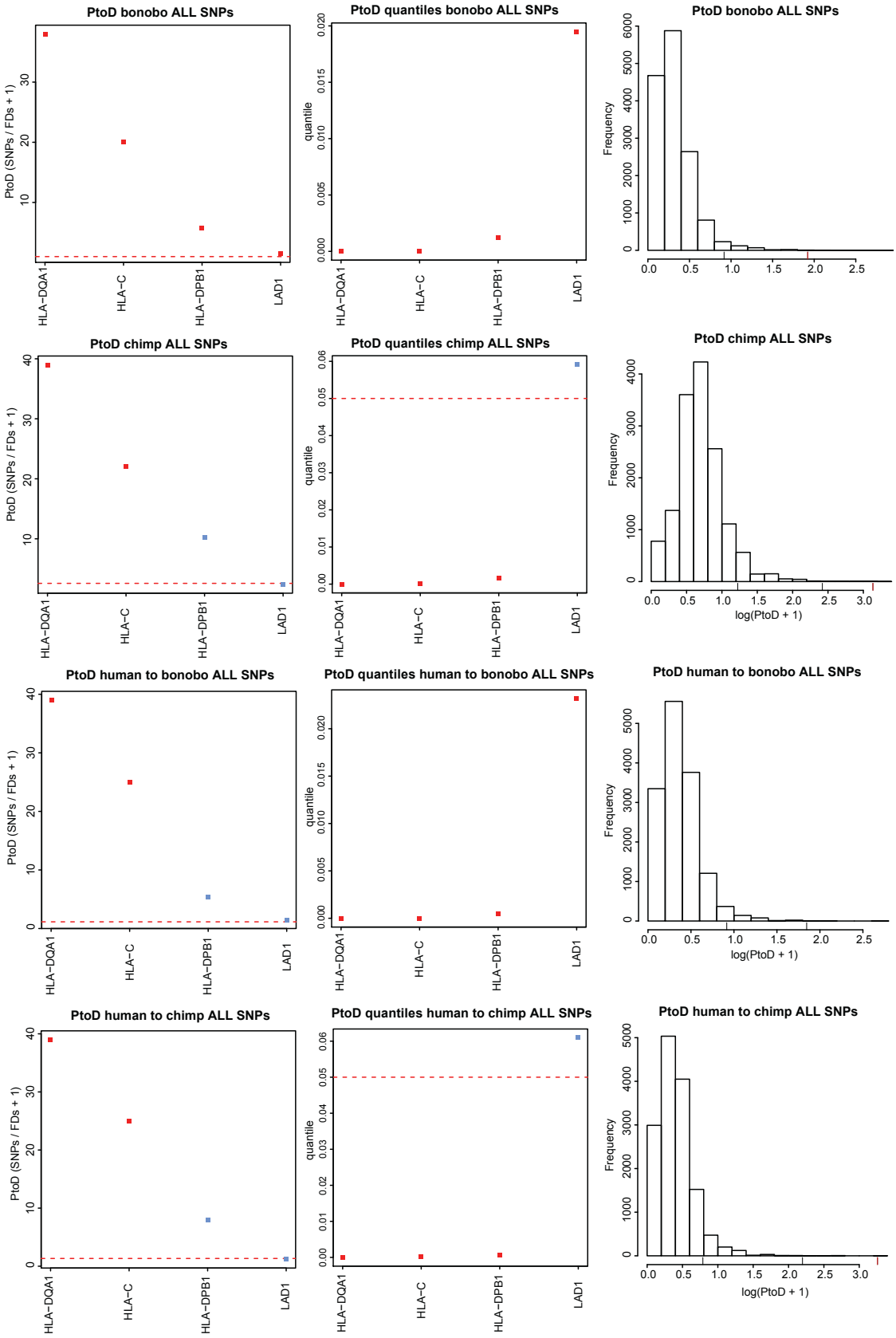
Comparable levels of diversity to HLA genes among candidates were only found in *LAD1*. The gene is also highly polymorphic compared to remainder of the genome, and lies in the upper tail of the empirical distribution in all species (significant at the 5% critical value in human for all comparisons – Table S4).

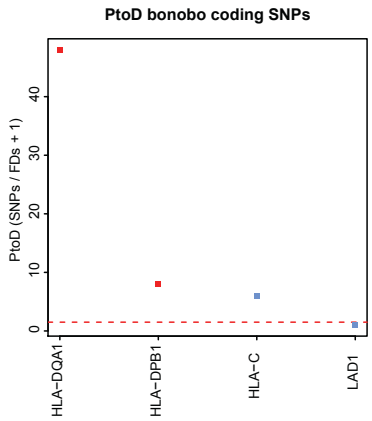
As for the other three genes (*LY9*, *SLCO1A2* and *OAS1*), the evidence for the action of long-term balancing selection as obtained from *PtoD* is rather poor: *LY9* shows average levels of diversity in all comparisons in all species with the only exception of the length of the allelic tree in humans, probably because there is another SNP in such short window (100 bp); *SLCO1A2* never shows significant excess of polymorphism; and *OAS1* shows significant excess polymorphism in chimpanzee (as was also shown by Ferguson et al. 2012) but not in bonobo and human. Because an unusually high level of polymorphism is a characteristic signature of the action of long-term balancing selection, these results (detailed in Tables 1 and S4) indicate that *LAD1*, *HLA-C*, *HLA-DQA1* and *HLA-DPB1* are the strongest candidate targets of long-standing balancing selection in our dataset.

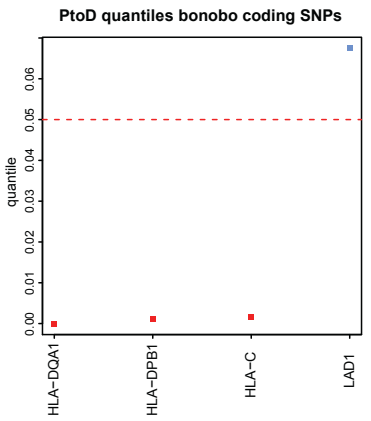
We then computed 2-tail Mann-Whitney U (MW-U) tests using R to assess whether the distribution of the average *PtoD* in the remaining genes (*HLA-C*, *HLA-DQA1*, *HLA-DPB1* and *LAD1*) was significantly greater than the distribution of control genes (Tables S2 and S3). After comparing the *PtoD* values in the two groups, we sequentially removed the top candidate (i.e. one gene each time) from the candidate's group and recalculated MW-U p-values maintaining the control group unaltered. This approach allowed us to control for the potential effects of a few known highly diverse candidates (i.e. *HLA* genes). We compared the

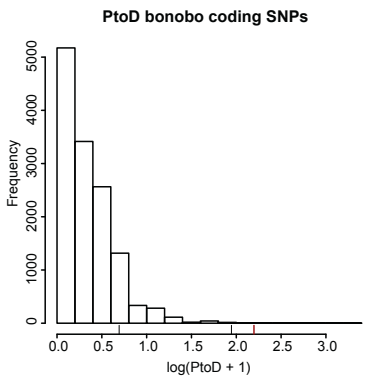
distribution of *PtoD* values for three of the five different sets mentioned above: 'ALL', 'coding', and '500bp'.

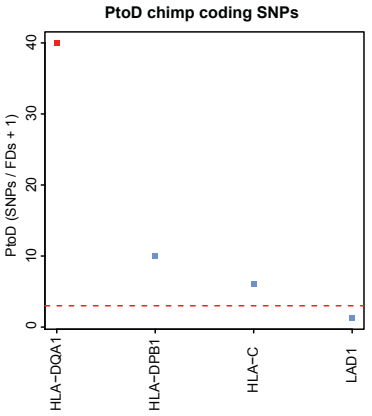
In all comparisons, the candidates' group was significantly more diverse than the control group (all $p < 2.1x10^{-3}$ – see Figures S6 and S7 and Table S2). In all species, and considering the three different comparisons, *HLA-DQA1* showed the highest *PtoD* values in all species and for all sets of comparisons, with *LAD1* being the least polymorphic of the group. Looking at the different species, chimpanzee showed the less – but still highly – significant increase in diversity for the candidates' group ($3.9x10^{-4} < p < 2.1x10^{-3}$) with 3 genes with $p < 0.05$ (most likely due to the higher effective population size of the central chimpanzees compared to human and bonobo), followed by human ($3.0x10^{-4} < p < 4.3x10^{-04}$) with 3-4 genes with $p < 0.05$, and bonobo ($3.1x10^{-4} < p < 4.9x10^{-4}$) with 3-4 genes with $P<0.05$ (table S3).

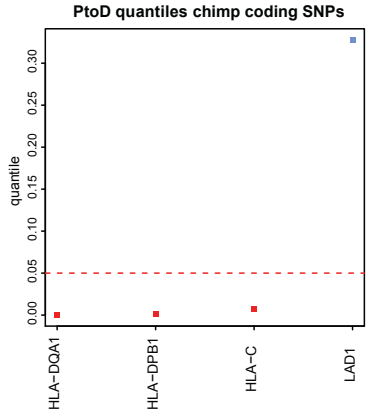

PtoD bonobo ALL SNPs
PtoD (SNPs / FDs + 1)
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD quantiles bonobo ALL SNPs
quantile
PtoD bonobo ALL SNPs
Frequency
PtoD chimp ALL SNPs
PtoD quantiles chimp ALL SNPs
PtoD chimp ALL SNPs
log(PtoD + 1)
PtoD human to bonobo ALL SNPs
PtoD quantiles human to bonobo ALL SNPs
PtoD human to bonobo ALL SNPs
PtoD human to chimp ALL SNPs
PtoD quantiles human to chimp ALL SNPs
PtoD human to chimp ALL SNPs

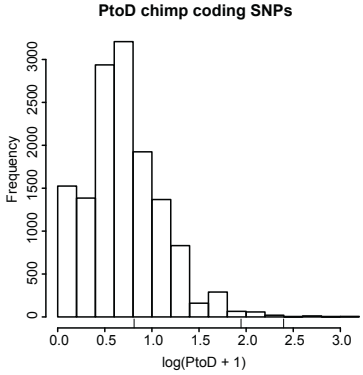

PtoD bonobo coding SNPs
PtoD (SNPs / FDs + 1)
0
10
20
30
40
HLA−DQA1
HLA−DPB1
HLA−C
LAD1

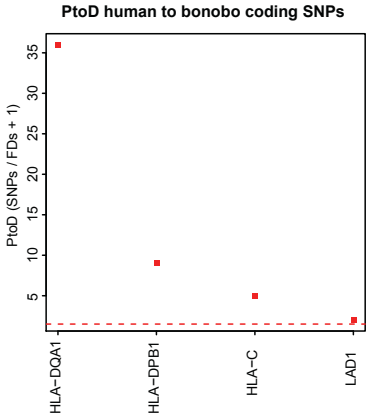

PtoD quantiles bonobo coding SNPs
quantile
0.00
0.01
0.02
0.03
0.04
0.05
0.06
HLA−DQA1
HLA−DPB1
HLA−C
LAD1

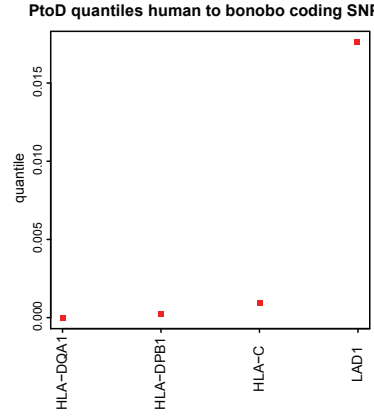

PtoD bonobo coding SNPs
Frequency
0
1000
2000
3000
4000
5000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)

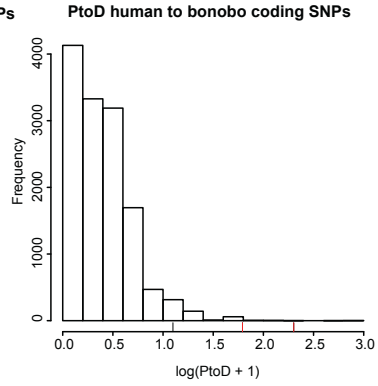

PtoD chimp coding SNPs
PtoD (SNPs / FDs + 1)
0
10
20
30
40
HLA−DQA1
HLA−DPB1
HLA−C
LAD1

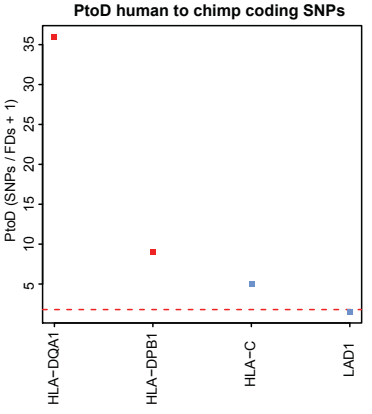

PtoD quantiles chimp coding SNPs
quantile
0.00
0.05
0.10
0.15
0.20
0.25
0.30
HLA−DQA1
HLA−DPB1
HLA−C
LAD1

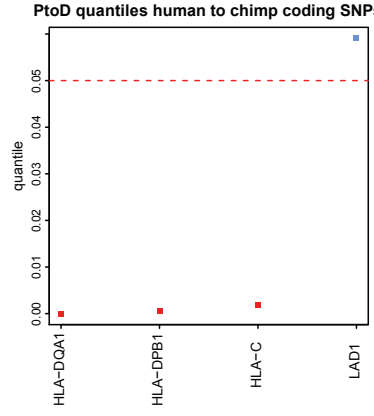

PtoD chimp coding SNPs
Frequency
0
500
1000
1500
2000
2500
3000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)

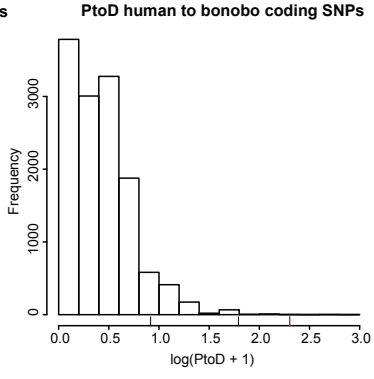

PtoD human to bonobo coding SNPs
PtoD (SNPs / FDs + 1)
5
10
15
20
25
30
35
HLA−DQA1
HLA−DPB1
HLA−C
LAD1

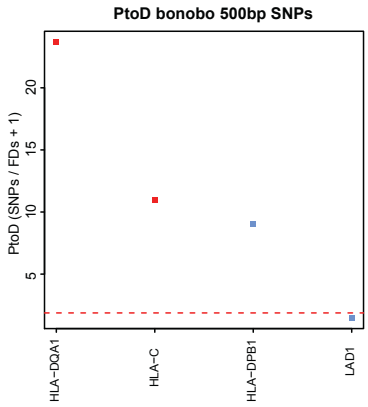

PtoD quantiles human to bonobo coding SNPs
quantile
0.000
0.005
0.010
0.015
HLA−DQA1
HLA−DPB1
HLA−C
LAD1

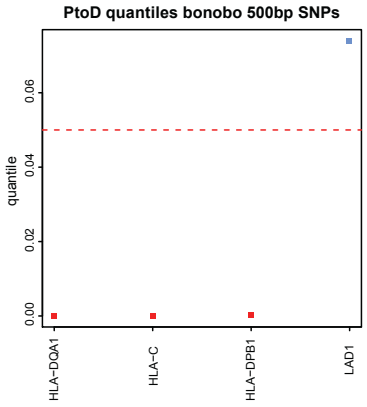

PtoD human to bonobo coding SNPs
Frequency
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)

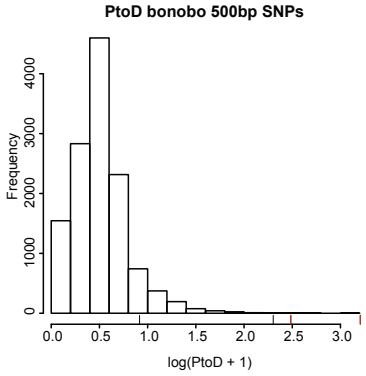

PtoD human to chimp coding SNPs
PtoD (SNPs / FDs + 1)
5
10
15
20
25
30
35
HLA−DQA1
HLA−DPB1
HLA−C
LAD1

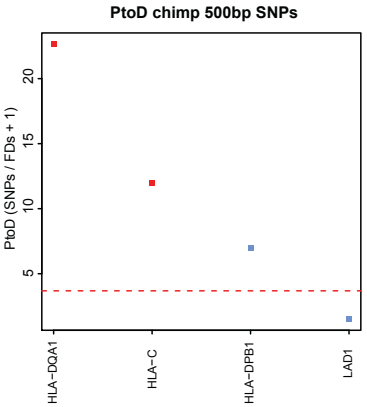

PtoD quantiles human to chimp coding SNPs
quantile
0.00
0.01
0.02
0.03
0.04
0.05
HLA−DQA1
HLA−DPB1
HLA−C
LAD1

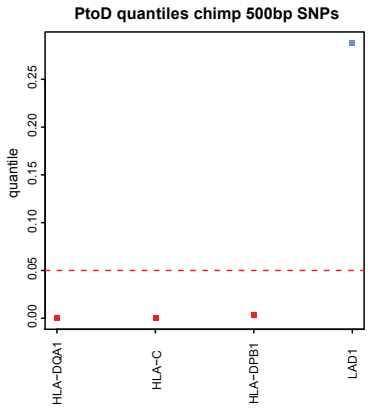

PtoD human to bonobo coding SNPs
Frequency
0
1000
2000
3000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)

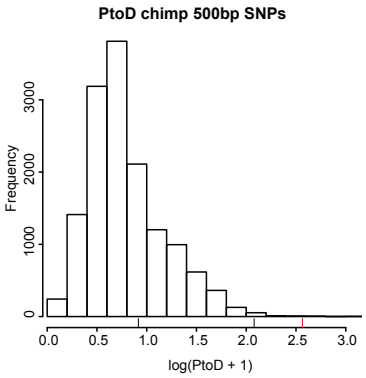

PtoD bonobo 500bp SNPs
PtoD (SNPs / FDs + 1)
HLA−DQA1
HLA−C
HLA−DPB1
LAD1

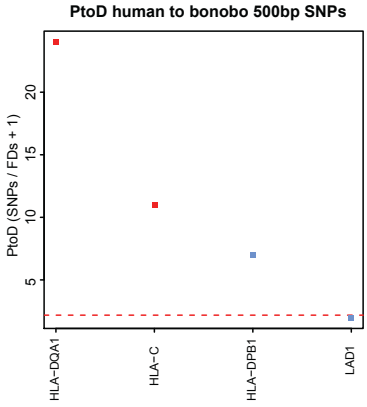

PtoD quantiles bonobo 500bp SNPs
quantile
HLA−DQA1
HLA−C
HLA−DPB1
LAD1

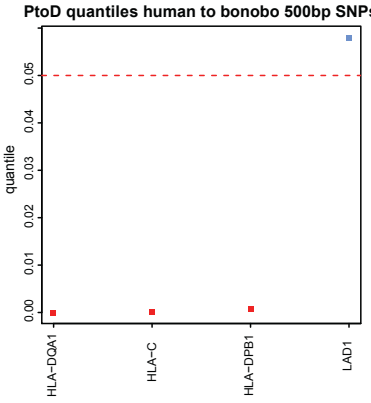

PtoD bonobo 500bp SNPs
Frequency
log(PtoD + 1)

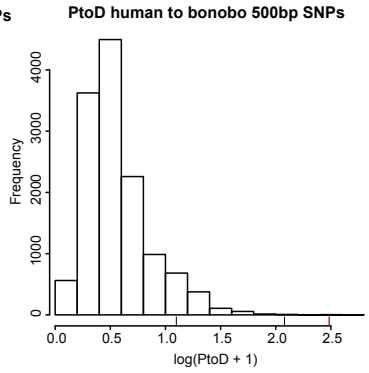

PtoD chimp 500bp SNPs
PtoD (SNPs / FDs + 1)
HLA−DQA1
HLA−C
HLA−DPB1
LAD1

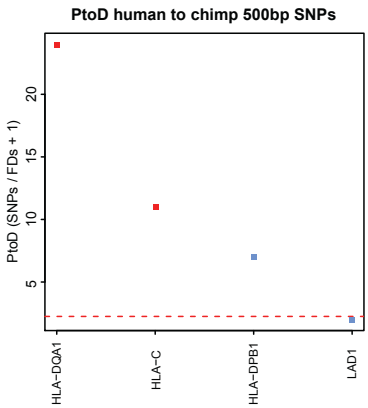

PtoD quantiles chimp 500bp SNPs
quantile
HLA−DQA1
HLA−C
HLA−DPB1
LAD1

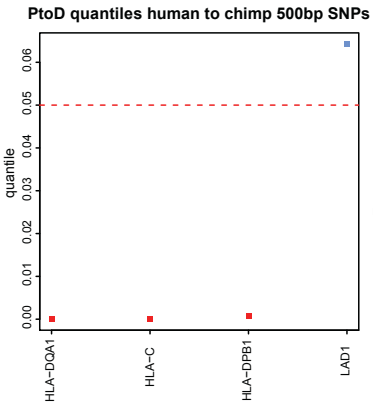

PtoD chimp 500bp SNPs
Frequency
log(PtoD + 1)


PtoD human to bonobo 500bp SNPs
PtoD (SNPs / FDs + 1)
HLA−DQA1
HLA−C
HLA−DPB1
LAD1


PtoD quantiles human to bonobo 500bp SNPs
quantile
HLA−DQA1
HLA−C
HLA−DPB1
LAD1


PtoD human to bonobo 500bp SNPs
Frequency
log(PtoD + 1)


PtoD human to chimp 500bp SNPs
PtoD (SNPs / FDs + 1)
HLA−DQA1
HLA−C
HLA−DPB1
LAD1


PtoD quantiles human to chimp 500bp SNPs
quantile
HLA−DQA1
HLA−C
HLA−DPB1
LAD1

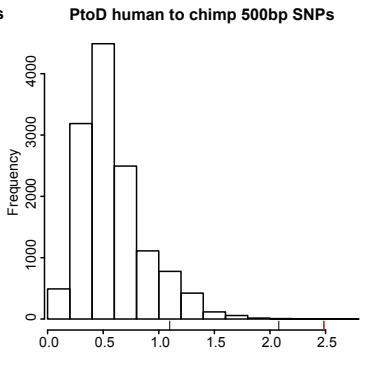

PtoD human to chimp 500bp SNPs
Frequency

Figure S6: PtoD ratios in candidate genes considering different sets of SNPs: 'ALL' includes all SNPs and FDs found in each gene; 'coding' represents variants found in the exons; and '500bp' represents the average PtoD values for 500bp (-/+ 250bp) windows surrounding shSNPs in each gene. Left plots show the actual *PtoD* ratios for the candidate genes in each species and for each set, separately. The quantile values for each gene (considering all targeted genes) are shown in the middle, whereas its distribution can be seen in the histogram on the right (red bars represent genes for which P<0.05).

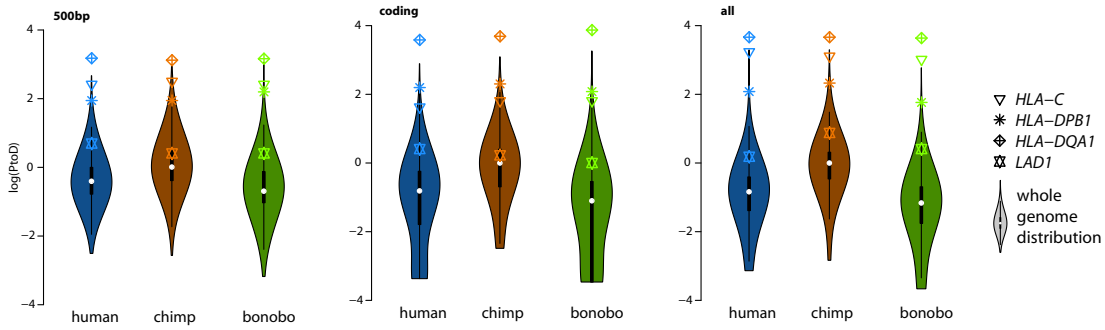


Figure S7: Violin plots of PtoD distributions of the controls (dark color) and the four genes (light color with symbols specified in the legend). The values are calculated: in 500 bp window centered on the SNP **(A)**; using only coding exonic regions **(B)**; and for the complete genes' sequence **(C)**. The plots were created using the R function 'vioplot' from 'vioplot' package (Hintze, Nelson 1998) with default parameters.

PtoD ALL													
rank	Gene	bonobo			chimp			human to bonobo			human to chimpanzee		
		PtoD	MW-U p-value	P	PtoD	MW-U p-value	P	PtoD	MW-U p-value	P	PtoD	MW-U p-value	P
1	*HLA-DQA1*	38.00	1.55E-03	0.000	39.00	2.01E-03	0.000	39.00	1.60E-03	0.000	39.00	2.02E-03	0.000
2	*HLA-C*	20.00	8.30E-03	0.000	22.00	1.08E-02	0.000	25.00	8.54E-03	0.000	25.00	1.08E-02	0.000
3	*HLA-DPB1*	5.83	4.86E-02	0.001	10.25	6.38E-02	0.002	5.33	5.03E-02	0.000	8.00	6.47E-02	0.001
4	*LAD1*	1.50	NA	0.019	2.40	NA	0.059	1.50	NA	0.023	1.20	NA	0.061
PtoD coding													
rank	Gene	bonobo			chimp			human to bonobo			human to chimpanzee		
		PtoD	MW-U p-value	P	PtoD	MW-U p-value	P	PtoD	MW-U p-value	P	PtoD	MW-U p-value	P
1	*HLA-DQA1*	48.00	2.55E-03	0.000	40.00	1.04E-02	0.000	36.00	1.55E-03	0.000	36.00	2.01E-03	0.000
2	*HLA-C*	8.00	1.40E-02	0.001	10.00	5.38E-02	0.002	9.00	8.37E-03	0.000	9.00	1.09E-02	0.001
3	*HLA-DPB1*	6.00	8.61E-02	0.002	6.00	2.83E-01	0.008	5.00	4.99E-02	0.001	5.00	6.57E-02	0.002
4	*LAD1*	1.00	NA	0.067	1.25	NA	0.328	2.00	NA	0.018	1.50	NA	0.059
PtoD 500bp													
rank	Gene	bonobo			chimp			human to bonobo			human to chimpanzee		
		PtoD	MW-U p-value	P	PtoD	MW-U p-value	P	PtoD	MW-U p-value	P	PtoD	MW-U p-value	P
1	*HLA-DQA1*	23.67	2.23E-03	0.000	22.67	8.03E-03	0.000	24.00	2.08E-03	0.000	24.00	2.17E-03	0.000
2	*HLA-C*	11.00	1.20E-02	0.000	12.00	4.22E-02	0.001	11.00	1.11E-02	0.000	11.00	1.17E-02	0.000
3	*HLA-DPB1*	9.00	7.28E-02	0.000	7.00	2.35E-01	0.003	7.00	6.65E-02	0.001	7.00	6.97E-02	0.001
4	*LAD1*	1.50	NA	0.074	1.50	NA	0.288	2.00	NA	0.058	2.00	NA	0.064

Table S2: PtoD values for the set of candidate genes. *MW-U P* is the recalculated Mann-Whitney U p-value (when comparing the candidate and the control sets) after removing the top-score gene from the candidate set. *P* is the percentile of each gene in the overall distribution.

PtoD			# genes	
	set	MW-U *P*	MW-U < 0.05	P <0.05
bonobo	ALL SNPs	3.05E-04	4	4
	coding SNPs	4.84E-04	2	3
	500bp	4.34E-04	2	3
chimpanzee	ALL SNPs	3.93E-04	2	3
	coding SNPs	2.05E-03	1	3
	500bp	1.59E-03	2	3
human to bonobo	ALL SNPs	3.14E-04	2	4
	coding SNPs	2.98E-04	4	4
	500bp	4.07E-04	2	3
human to chimpanzee	ALL SNPs	3.95E-04	2	3
	coding SNPs	3.88E-04	2	3
	500bp	4.25E-04	2	3

Table S3: MW-U p-values comparing PtoD in candidate and control genes (genomic distribution of diversity). The numbers of genes that are significant in the MW-U ranked test (MW-U<0.05), as well as the number of genes in the top 5% of the distribution of PtoD (P<0.05) are shown.

						PtoD															
						ALL				coding				500bp				length of allelic tree				ALL
shSNP	chr	position (hg19)	gene	type	tree (>0.9)	bonobo	chimp	hum-bon	hum-chimp	bonobo	chimp	hum-bon	hum-chimp	bonobo	chimp	hum-bon	hum-chimp	bonobo	chimp	hum-bon	hum-chimp	3spp
1	1	155169633	*THBS3*	Synonymous	100bp																	
2	1	160788067	*LY9*	Replacement	100bp	0.40 (0.344)	1.10 (0.385)	0.33 (0.576)	0.34 (0.614)	0.53 (0.268)	0.81 (0.582)	0.27 (0.634)	0.25 (0.665)	0.25 (0.941)	0.75 (0.811)	0.75 (0.471)	0.75 (0.523)	1.00 (0.163)	1.00 (0.299)	2.00 (0.033)	2.00 (0.039)	1.65 (0.426)
3	1	201355761	*LAD1*	Replacement	350bp	1.50 (0.019)	2.40 (0.059)	1.50 (0.023)	1.20 (0.061)	1.00 (0.068)	1.25 (0.332)	2.00 (0.018)	1.50 (0.061)	1.50 (0.069)	1.50 (0.317)	2.00 (0.024)	2.00 (0.033)	3.00 (0.024)	3.00 (0.074)	3.00 (0.028)	3.00 (0.034)	4.2 (0.028)
4	5	34826545	*RAI14*	Synonymous	950bp																	
5	8	22995487	*TNFRSF10D*	Replacement	100bp																	
6	8	23003292	*TNFRSF10D*	Replacement	150bp																	
7	9	80040566	*GNA14*	Synonymous	150bp																	
8	9	114490308	*C9orf84*	Replacement	150bp																	
9	12	21453466	*SLCO1A2*	Synonymous	1000bp	0.94 (0.074)	1.58 (0.176)	0.59 (0.258)	0.61 (0.293)	0.73 (0.190)	2.09 (0.121)	0.55 (0.337)	0.55 (0.378)	1.00 (0.190)	2.00 (0.144)	1.00 (0.231)	1.00 (0.273)	1.00 (0.168)	2.00 (0.214)	1.00 (0.210)	1.00 (0.241)	2.625 (0.117)
10	12	113354384	*OAS1*	Replacement	250bp	0.48 (0.276)	5.29 (0.009)	0.38 (0.524)	1.14 (0.068)	0.43 (0.374)	14.00 (0.001)	0.21 (0.692)	1.50 (0.061)	0.25 (0.941)	9.00 (0.002)	0.25 (0.955)	1.00 (0.273)	1.00 (0.195)	6.00 (0.012)	1.00 (0.229)	1.00 (0.247)	2.33 (0.164)
11	13	75915285	*TBC1D4*	Synonymous	100bp																	
12	15	33855175	*RYR3*	Synonymous	300bp																	
13	17	31260255	*TMEM98*	Synonymous	100bp																	
14	6	31237124	*HLA-C*	Replacement	150bp	20.00 (0.000)	22.00 (0.000)	25.00 (0.000)	25.00 (0.000)	6.00 (0.002)	6.00 (0.008)	5.00 (0.001)	5.00 (0.002)	11.00 (0.000)	12.00 (0.001)	11.00 (0.000)	11.00 (0.000)	2.00 (0.048)	3.00 (0.054)	2.00 (0.051)	2.00 (0.060)	38.00 (0.000)
15	6	32609097	*HLA-DQA1*	Synonymous	100bp													5.00 (0.005)	7.00 (0.003)	11.00 (0.000)	11.00 (0.000)	
16	6	32609105	*HLA-DQA1*	Replacement	250bp	38.00 (0.000)	39.00 (0.000)	39.00 (0.000)	39.00 (0.000)	32.00 (0.000)	32.00 (0.000)	27.00 (0.000)	27.00 (0.000)	23.67 (0.000)	22.67 (0.000)	24.00 (0.000)	24.00 (0.000)	13.00 (0.000)	14.00 (0.002)	19.00 (0.000)	19.00 (0.000)	60.00 (0.000)
17	6	32609271	*HLA-DQA1*	Replacement	750bp													24.00 (0.000)	23.00 (0.000)	25.00 (0.000)	25.00 (0.000)	
18	6	33052736	*HLA-DPB1*	Replacement	1000bp													11.00 (0.001)	10.00 (0.005)	8.00 (0.004)	8.00 (0.008)	
19	6	33052743	*HLA-DPB1*	Synonymous	1000bp	5.83 (0.001)	10.25 (0.002)	5.33 (0.000)	8.00 (0.001)	8.00 (0.001)	10.00 (0.002)	9.00 (0.000)	9.00 (0.001)	9.00 (0.001)	7.00 (0.003)	7.00 (0.000)	7.00 (0.000)	11.00 (0.001)	10.00 (0.005)	8.00 (0.004)	8.00 (0.008)	9.83 (0.002)
20	6	33052768	*HLA-DPB1*	Synonymous	1000bp													11.00 (0.001)	10.00 (0.005)	8.00 (0.004)	8.00 (0.008)	

						info sites (length of allelic tree)												
						bonobo		chimp		hum-bon		hum-chimp		MAF				
shSNP	chr	position (hg19)	gene	type	tree (>0.9)	SNPs	FDs	SNPs	FDs	SNPs	FDs	SNPs	FDs	human	chimp	bonobo	BLAT	Note
1	1	155169633	*THBS3*	Synonymous	100bp									0.025	0.100	0.225	duplication HC	
2	1	160788067	*LY9*	Replacement	100bp	1 (1.000)	0 (0.574)	1 (1.000)	0 (0.605)	2 (0.264)	0 (0.549)	2 (0.264)	0 (0.587)	0.100	0.300	0.125	ok	
3	1	201355761	*LAD1*	Replacement	350bp	3 (0.194)	0 (0.267)	3 (0.462)	0 (0.300)	3 (0.266)	0 (0.258)	3 (0.266)	0 (0.305)	0.450	0.325	0.225	ok	
4	5	34826545	*RAI14*	Synonymous	950bp									0.025	0.100	0.075	duplication HC	
5	8	22995487	*TNFRSF10D*	Replacement	100bp									0.075	0.050	0.025	duplication HC	
6	8	23003292	*TNFRSF10D*	Replacement	150bp									0.125	0.400	0.250	duplication C	
7	9	80040566	*GNA14*	Synonymous	150bp									0.075	0.275	0.125	duplication H	overlaps
8	9	114490308	*C9orf84*	Replacement	150bp									0.300	0.025	0.100	duplication C	overlaps
9	12	21453466	*SLCO1A2*	Synonymous	1000bp	2 (0.630)	1 (0.307)	4 (0.520)	1 (0.421)	2 (0.711)	1 (0.330)	2 (0.711)	1 (0.382)	0.025	0.350	0.050	duplication HC	no overlap
10	12	113354384	*OAS1*	Replacement	250bp	1 (1.000)	0 (0.351)	6 (0.059)	0 (0.351)	1 (1.000)	0 (0.330)	1 (1.000)	0 (0.365)	0.025	0.350	0.025	ok	
11	13	75915285	*TBC1D4*	Synonymous	100bp									0.025	0.025	0.050	duplication HC	
12	15	33855175	*RYR3*	Synonymous	300bp									0.050	0.025	0.025	duplication H	overlaps
13	17	31260255	*TMEM98*	Synonymous	100bp									0.025	0.425	0.100	duplication HC	
14	6	31237124	*HLA-C*	Replacement	150bp	2 (0.302)	0 (0.472)	3 (0.240)	0 (0.486)	2 (0.366)	0 (0.470)	2 (0.366)	0 (0.512)	0.225	0.225	0.225	na	
15	6	32609097	*HLA-DQA1*	Synonymous	100bp	5 (0.020)	0 (0.574)	7 (0.005)	0 (0.605)	11 (0.000)	0 (0.549)	11 (0.000)	0 (0.587)	0.200	0.400	0.050	na	no-CpG
16	6	32609105	*HLA-DQA1*	Replacement	250bp	13 (0.002)	0 (0.351)	14 (0.002)	0 (0.351)	19 (0.000)	0 (0.330)	19 (0.000)	0 (0.365)	0.500	0.350	0.050	na	no-CpG
17	6	32609271	*HLA-DQA1*	Replacement	750bp	24 (0.003)	0 (0.136)	23 (0.001)	0 (0.152)	25 (0.000)	0 (0.127)	25 (0.000)	0 (0.162)	0.475	0.400	0.050	na	
18	6	33052736	*HLA-DPB1*	Replacement	1000bp	11 (0.010)	0 (0.124)	10 (0.071)	0 (0.161)	8 (0.036)	0 (0.130)	8 (0.036)	0 (0.160)	0.300	0.425	0.325	na	
19	6	33052743	*HLA-DPB1*	Synonymous	1000bp	11 (0.010)	0 (0.124)	10 (0.071)	0 (0.161)	8 (0.036)	0 (0.130)	8 (0.036)	0 (0.160)	0.300	0.450	0.350	na	
20	6	33052768	*HLA-DPB1*	Synonymous	1000bp	11 (0.010)	0 (0.124)	10 (0.071)	0 (0.161)	8 (0.036)	0 (0.130)	8 (0.036)	0 (0.160)	0.300	0.475	0.350	na	

Table S4: Summary table of the 20 shSNPs that passed the shSNP specific filtering. Polymorphism to divergence ratios, number of informative sites in the length of the allelic tree and MAF are also shown. 'Note' includes information on whether the duplicated region overlaps with the candidate trSNP, and whether the trSNP lies in a CpG site. Green cells highlight significant tests and red entries highlight trSNPs that show evidence of duplication (additional tests were not performed). H – human; C – chimpanzee.

V – Tajima's D

A classical test to detect departures from neutrality in the genome is Tajima's D (Tajima 1989). Particularly, a positive value of Tajima's D – caused by an excess of intermediate-frequency alleles – is classically considered as a signature of balancing selection in the absence of population substructure. We calculated Tajima's D for the set of trSNP-containing genes and a set of control genes (considering all genes with at least six polymorphic sites, which is the minimum number of SNPs in the four genes containing a trSNP). Because we targeted the exons, Tajima's D for the control set are expected to have on average slightly negative values due to the action of purifying selection, and we observe that shift (Figure S8). On the contrary, all candidate genes have positive Tajima's D, with the exception of *LAD1* in bonobo. These values, however, are not all significantly higher when compared with the control genes (5% upper tail cutoff, Figure S8). We note that the power of this test is hampered by the limited number of SNPs in the coding regions of genes.

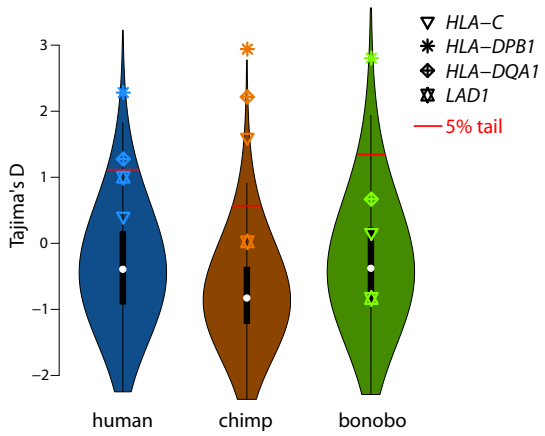


Figure S8 – Violin plots representing the distribution of Tajima's D in control genes (dark color) and candidate genes (represented by symbols) in all species. The plots have been generated with the 'vioplot' function in R (Hintze and Nelson 1998) using default values. The red line represents the 5% upper tail boundary of the distribution for each species.

VI – Comparison with available datasets

In order to compare our results to previously published studies, we investigated whether additional shSNPs in candidate genes (that we might have missed) were present in a whole-genome dataset consisting of several sequenced individuals from different great ape species (Prado-Martínez et al. 2013). We also verified whether some of the trSNPs uncovered in our study were found in a genome-wide scan for long-term balancing selection in humans and chimpanzees (Leffler et al. 2013).

If we compare the trSNPs found in this study to the dataset of Prado-Martinez et al. (2013), out of the 8 trSNPs uncovered in our study (in four genes), six are also shared between humans, chimpanzees and bonobos in that dataset, including rs12088790. Moreover, we identify one additional shSNP downstream of *LAD1* (chr1:201349024) that is also present in all three species (Table S8). This SNP, which was previously described in humans (rs12035254), is located far from rs12088790 (~6 kbp) and thus cannot explain the signatures of balancing selection in the short region containing rs12088790.

However, the picture looks different when we attempt to retrieve our eight trSNP from the set of human-chimpanzee trSNPs that were identified as part of short trans-species haplotypes in Leffler et al. (2013), as we can detect none. This is probably due to the different strategies adopted in the studies, as we focused our analysis on shared polymorphism on the coding sequences of the genome, while Leffler et al. (2013) focused on shared haplotypes (with at least 2 SNPs with significant linkage disequilibrium), which happened to be largely non-coding. We though also searched for the presence of our trSNPs in a list of single coding shSNPs provided by Leffler et

al. (2013), and retrieved none. Although perhaps surprising, the lack of correspondence might be explained by a number of differences between the two studies regarding samples and coverage depth. For example, we analyze 20 individuals per species with an average coverage of ~18X in all species; Leffler et al. (2013) analyzed a genome-wide dataset with only moderate coverage (~9X for the chimpanzees and 3.4X for the human samples), with smaller chimpanzee sample size (10 individuals) and much larger human sample size (59 individuals) than our dataset. In addition, the two studies analyzed different chimpanzee subspecies (*Pan troglodytes troglodytes* vs *Pan troglodytes verus*). Nevertheless, we note that Leffler et al. (2013) uncovered human-chimpanzee shSNPs in the two HLA genes where we identify trSNPs, although the SNPs identified are different (4 shSNPs in *HLA-DQA1*, 3 shSNPs in *HLA-DPB1*).

VII – Supplementary Tables

human	chimpanzee	bonobo
NA18501	Agnagui	Api
NA18504	Bailele	Bandundu
NA18505	Bayokele	BilliL
NA18508	Bimangou	Boende
NA18516	Botsomi	Bolobo
NA18522	Casimir	Fizi
NA18523	Castro	Isiro
NA18853	Chinoc	Keza
NA18856	ClaraT	Kikwit
NA18858	Dzeke	Kisantu
NA18861	Elikia	Kubulu
NA18870	FanTuek	Likasi
NA18871	Gao	Lipopo
NA18912	Golfi	Lodja
NA19093	GrandMaitre	Lomami
NA19102	Imphondo	MalouL
NA19137	Loufoumbou	Matadi
NA19138	Lufino	Max
NA19238	Marcelle	Semendwa
NA19239	Moka	Tshilomba

Table S5: The 20 humans, 20 chimpanzees and 20 bonobos used in this study.

POP	*p*
ASW	0.0008
LWK	0.0019
YRI	0.0262
CEU	0.9999
FIN	0.4318
GBR	0.9763
TSI	0.9768
CHB	0.3338
CHS	0.7786
JPT	0.2588
MXL	0.9632
CLM	0.8802
PUR	0.1560

Table S6: P-values of the MWU test between the MAF of *LAD1* and that of the entire chromosome 1 in the 1000Genomes (Abecasis et al. 2012) populations. Values closer to 0 and 1 indicate shift towards intermediate- and low-frequency alleles, respectively. We filtered the 1000Genomes data by considering only SNPs that: 1) are in the 50mer mappability track (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability); 2) are not in the Tandem Repeat Finder; 3) are not in annotated segmental duplications (Cheng et al. 2005; Alkan et al. 2009; Prüfer et al. 2012); and 4) are perfectly aligned to PanTro2 genome.

Mutations	human		chimpanzee		bonobo	
	all	shGenes	all	shGenes	all	shGenes
Synonymous (S)	18,955	21	43,023	21	15,549	19
Non-Synonymous (NS)	18,208	26	36,151	31	15,079	30
NS/S ratio	0.96	1.23	0.84	1.48	0.97	1.44

Table S7: Number of synonymous and non-synonymous SNPs for each species using all coding SNPs and those falling within the four candidate genes 'shGenes'. χ^2 test of differences between 'all' and 'shGenes' for each species are all not significant.

chromosome	position (hg19)	type
1	201355761	non-synonymous
1	201349024	intronic

Table S8: shSNPs uncovered in the gene *LAD1* across the three species in Prado-Martínez (2013).

YRI vs	F_{ST} (rs12088790)	*p*	$\overline{F}_{ST}$
CEU	0.355	0.114	0.148
FIN	0.355	0.118	0.150
GBR	0.334	0.144	0.149
TSI	0.238	0.263	0.146
CHB	0.293	0.264	0.171
CHS	0.256	0.310	0.172
JPT	0.334	0.189	0.171
MXL	0.334	0.108	0.136
CLM	0.256	0.175	0.123

Table S9: F_{ST} between African (YRI) and non-African populations in the 1000 genomes (Abecasis et al. 2012) for rs12088790. P-values (*p*) were obtained by comparing the F_{ST} of rs12088790 to the F_{ST} distribution obtained for alleles at intermediate frequency in YRI (0.30 < MAF < 0.5), which are similar to the frequency of rs12088790 (DAF=0.38). The mean $\overline{F}_{ST}$ of these distributions of genome-wide F_{ST} is also reported.

5. Signatures of long-term balancing selection in the genomes of great apes

João C. Teixeira[1], Joshua Schmidt[1], Cesare de Filippo[1], Bárbara Bitarello[2] and Aida M. Andrés[1]

[1]*Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6 04103 Leipzig, Germany*

[2]*Department of Genetics and Evolutionary Biology, University of São Paulo, Rua do Matão, 277, São Paulo, Brazil*

(in preparation for publication)

5.1 Abstract

Balancing selection maintains advantageous genetic diversity in populations and can shape the evolution of selected loci for millions of years. In fact, classical examples of long-term balancing selection in great apes include trans-species polymorphisms (trSNPs) at the major-histocompatibility complex (MHC) and the ABO locus. These polymorphisms unveil remarkable, extreme signatures of balancing selection and are expected to represent a minor fraction of the loci evolving under similar selective pressures in different species. Our study aims to determine to what extent long-term balancing selection is shared among great ape species, outside of the extreme examples provided by trans-species polymorphisms. We analyzed genome-wide population samples from nine great ape subspecies and identified candidate targets of long-term balancing selection in each population by implementing a statistic (NCD2) that has high power to detect departures from the distribution of neutral allele frequencies. We provide evidence for a significant excess of shared targets of balancing selection across all species, which cannot be accounted for by shared ancestry between them. Remarkably, this pattern is also present among species that diverged millions of years ago. Furthermore, we show that balancing selection mostly affects genes and has, at the same time, influenced the evolution of regulatory regions of the genome. Overall, this study provides a concerted catalog of signatures of balancing selection on all of major great ape clades, and demonstrates that its targets are shared between species aside classical examples provided by the trSNPs.

5.2 Introduction

Balancing selection maintains advantageous genetic diversity in populations through a variety of mechanisms (Andrés 2011; Key et al. 2014) including overdominance or heterozygous advantage (Allison 1956; Pasvol et al. 1978), frequency dependent selection favoring rare alleles (Wright 1939; Gigord et al. 2001), temporal or spatial variation in selective pressures (Gillespie 1978; Muehlenbachs et al. 2008), or pleiotropy (Gendzekhadze et al. 2009).

Because selection actively maintains balanced polymorphisms, they can segregate in populations for far longer than neutral variation. This results in deeper local genealogies (older time to the most recent ancestor, TMRCA) where neutral variants arising from new mutations and that are linked to selected polymorphisms start to accumulate (Charlesworth et al. 1997, Charlesworth 2006, Andrés 2011, Key et al. 2014). Furthermore, both selected and linked neutral polymorphisms segregate at frequencies close to the frequency-equilibrium, which is the allele frequency that maximizes fitness in the population (Hudson and Kaplan 1988, Takahata and Nei 1990, Charlesworth et al. 19997, Charlesworth 2006). The longer selected polymorphisms and linked neutral variation are maintained in a particular population, the more the action of recombination through time restricts the length of regions exhibiting older TMRCA (Wiuf et al. 2004, Chalesworth 2006, Ségurel 2013, Teixeira et al. 2015).

Current knowledge on the role played by balancing selection on the evolution of primate genomes is largely limited to studies performed in humans, first by using candidate gene approach strategies, and more recently using genome-wide data across different human populations (Kroyman and Mitchell-Olds 2005, Hughes and Nei

1989, Bamshad et al. 2002, Wooding et al. 2005, Andrés et al. 2009, DeGiorgio et al. 2014). Among the targets of balancing selection in humans, arguably the most renowned example is the major histocompatibility locus (MHC), which contains extremely high levels of genetic diversity that likely improves the effectiveness of immune response via antigen presentation plasticity (Harding and Geuze 1993, Germain 1994, Hughes and Yeager 1998, Charlesworth 2006). Another well established case of balancing selection in humans is the β-globin gene in populations where malaria is endemic, on which the sickle cell allele is maintained at relatively high frequencies (~10%) because heterozygous individuals are more resistant to infection by *Plasmodium falciparum* than homozygous healthy individuals (Allison 1956). Apart from these classic textbook examples, other instances of balancing selection in humans have been inferred that show that selection mainly affects the evolution of immune-related genes (Bamshad et al. 2002, Asthana et al. 2005, Prugnolle et al. 2005, Bubb et al. 2006, Cagliani et al. 2008, Ferrer-Admetlla et al. 2008, Fumagalli et al. 2009, Andrés et al. 2009).

In contrast to the progress made in understanding balancing selection in humans, little is known about this particular mode of selection in our closest living relatives, the non-human great apes. One well-characterized exception is the antiviral gene *OAS1*, which has been shown to show signatures of balancing selection in chimpanzee (Ferguson et al. 2012). In fact, most studies have used other species to identify targets of balancing selection in humans using trans-species polymorphisms (trSNPs) between humans and chimpanzees. When balancing selection acts continuously for millions of years, selected polymorphisms (and neutral ones tightly linked to them) are able to survive the split of species, resulting in trSNPs, which undoubtedly represent hallmarks of balancing selection (Klein et al. 1993, Asthana et al. 2005,

Cagliani et al. 2010, Ségurél et al. 2012, Leffler et al. 2013, Teixeira et al. 2015). The presence of trSNPs demonstrates not only that balancing selection can act for long evolutionary periods but also that its targets are most likely shared across different species. Nevertheless, evidence has so far shown that trSNPs are rare and therefore likely to represent only but a few of all the targets of balancing selection in the genome.

A study aiming to understand the extent to which balancing selection shaped the genomes of non-human great apes is essential to understand how balancing selection has affected – similarly or differently – the evolution of the genomes of different great apes species. At the same time, it opens the possibility to address questions of both functional and evolutionary importance. Which biological functions have evolved under balancing selection in the great apes? Is selection on all of these functions shared across ape species? If not, which are conserved and which are different?

The level of sharing of targets of balancing selection across species is a particularly interesting question. trSNPs are rare because they require virtually constant selective pressure, in the different species, for millions of years. But even if an ancestral balanced polymorphism is lost at a given moment in time (due to demographic or selective changes) genetic diversity may be advantageous for the same loci in different species.

Here, we take advantage of the recent availability of genome-wide data for all major great ape species (Prado-Martínez et al. 2013) and introduce a genome-wide analysis to identify targets of balancing selection in nine great ape species using the "Non-Central Deviation" (NCD2) (Bitarello et al. in prep) statistic, which was specifically designed for detecting targets of balancing selection in the genome. We show that candidate targets of balancing selection significantly overlap with genic regions of the

genome and are, to at least some extent, also involved in the regulation of gene expression. Furthermore, we provide evidence for a significant excess of targets that are shared by species that diverged millions of years in the past and that this pattern cannot be explained under neutrality. Finally, we uncovered five genes that show signatures of balancing selection in the exact same region in all nine species: *HLA-DRB1*, *HLA-DRB5*, *HLA-DQB1*, *ATN1* and *GARNL4*.

5.3 Results

5.3.1 Power analysis

We investigate the power of NCD2 to detect regions evolving under balancing selection by running simulations under realistic demographic scenarios for the great ape population history (based on demographic inferences in Prado-Martínez et al. 2013). We used the coalescent simulation software *msms* (Ewing and Hermison 2010) and run two separate sets of simulations per species: one set of neutral simulations and one set including balancing selection starting 3.5 million years in the past. We performed power analysis by comparing simulations under neutrality and under balancing selection, and present the results by means of a receiver operating characteristic (ROC) curve (Figure 1).

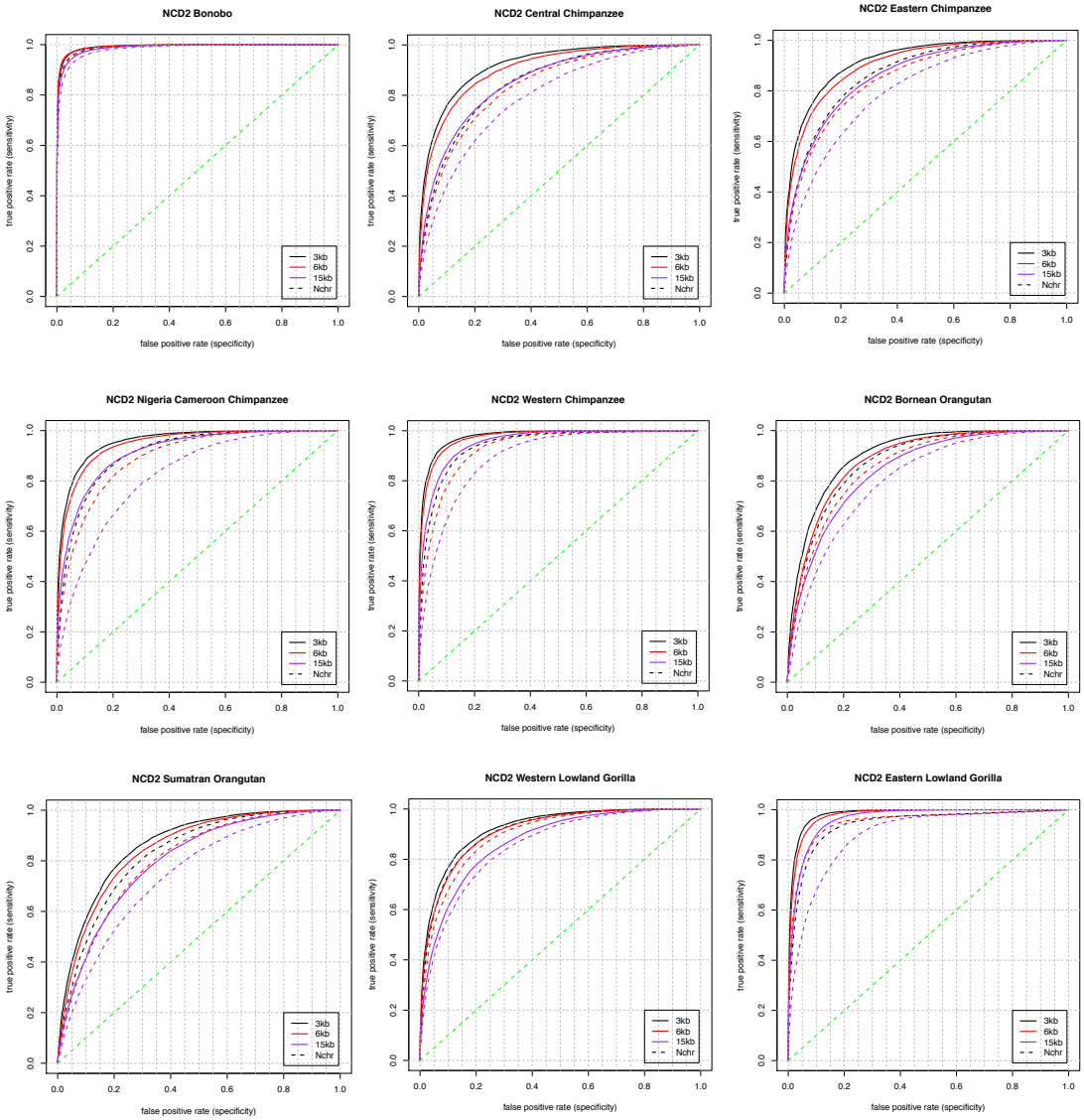


Figure 1: ROC curves showing the power of NCV to detect balancing selection in simulated data from different great ape species.

We start by comparing power across different window sizes (the length of the genomic region analyzed). In all species, NCD2 performs best when the size of the window is 3kb, when compared with larger window sizes. The only exception is eastern lowland gorilla, which has the lowest sample size (N=3) and shows comparable power for both 3kb and 6kb windows. These results are in agreement with expectations for long-term balancing selection, as the long-term action of

recombination disrupts the local signatures of selection and restricts them to a very small genomic region around the site under balancing selection. As expected, NCD2 performance is dependent on the sample size, with higher overall power when considering 50 chromosomes per species (dashed lines) than the actual sample size available from Prado-Martínez et al. (2013) (solid lines), simply because estimating allele frequencies from a small sample size is noisier than from a large sample size. Power is also particularly sensitive to the effective population size (Ne) of the species. Namely, NCD2 is particularly powerful at detecting balancing selection in species with low estimated long-term Ne, such as eastern lowland gorilla (Ne=2,000; true positive rate (TPR)>0.70 for false discovery rate (FDR)=0.05), western chimpanzee (Ne = 5,000; TPR>0.6 for FDR=0.05) and bonobo (Ne = 5,000; TPR>0.9 for FDR=0.05). The power of NCD2 is lower in uncovering targets of balancing selection in species with the highest estimated long-term Ne, as are the cases of central chimpanzee (Ne = 30,000; TPR>0.55 for FDR=0.05) or Sumatran orangutan (Ne = 17,000; TPR>0.4 for FDR=0.05). This in agreement with the fact that species with lower Ne have an overall depletion of genetic diversity at neutral loci, with most sites segregating at low frequencies, which enables NCD power to detect regions evolving under long-term balancing selection in these species.

Overall, our results indicate that NCD2 has considerably high power to detect the type of balancing selection we simulated, in all great ape species.

5.3.2 Uncovering targets of balancing selection

As mentioned in *Methods* (see below), NCD2 is expected to show higher variance in windows with low number of informative sites (*IS*), which can directly impact the detection of balancing selection targets because windows with lower number of *IS* are

likely to reach, under neutrality, more extreme NCD2 values than those with high numbers of *IS*, and thus make it more likely for a windows with low number of *IS* to be included within candidate targets. We investigated, in the genomic data, how the variance in NCD2 estimates changes with the number of *IS* per window by excluding the windows with the lowest number of *IS* (in increments from quantile 5% to quantile 30%) and calculating NCD2 variance in the remaining windows. At the same time, we considered the proportion of windows lost on each filtering step (again from quantile 5% to quantile 30%, Supplementary Table S1).

We show that variance in NCD2 is highest when considering all scanned windows and that it drops when excluding windows with the lowest number of *IS* (Supplementary Table S1), with the strongest effect observed when excluding windows in the 5% lowest quantile of the distribution of number of IS (Figure 2 and Supplementary Table S2).

Even though further filtering the data (for 10%, 15%, 20%, 25% and 30% quantiles of *IS*) decreases, slightly, NCD2's variance estimates, the reduction in variance is much weaker than for the 5% quantile (Supplementary Table S2). Furthermore, using a 5% quantile filtering obviously retains a higher proportion of windows than using other quantile for filtering, which prevents too much data loss (Supplementary Table S1).

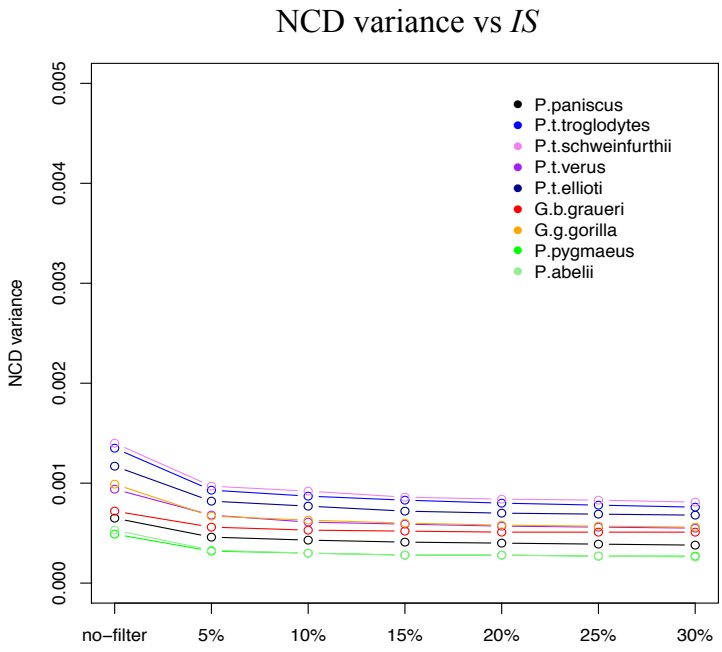


Figure 2: NCD2 variance estimates across the different species before and after filtering based on quantiles of *IS*

Hence, we proceed with our analysis by excluding windows with low number of *IS* as defined by the 5% quantile values on the number of *IS*. The total number of windows considered for further analysis is shown in Supplementary Table S3, with a minimum of 8 *IS* per window (Supplementary Table S1).

Finally, and in order to further account for possible differences in NCD2 variance across windows with different number of *IS*, we normalized NCD2 by calculating Z_{tf} values, a strategy proposed by Bitarello et al. (in prep) – see methods. This strategy minimizes biases in the identification of targets of balancing selection in the genome that arise from the different number of *IS* in different windows.

We identify putative targets of balancing selection by focusing on the lower-tails of the empirical distribution of Z_{tf} values ($P<0.001$ or $P<0.005$) in each species. This empirical outlier approach is based on the reasonable assumption that the tails of the empirical distribution of a statistic sensitive to balancing selection will be strongly

enriched in true targets of balancing selection. We focus on the very far tail of the distribution (0.1% to 0.5%) where (given the properties of NCD2) this is most likely to be true. In fact, this tails contains many of the few well-known targets of balancing selection in these species (e.g. MHC genes, see below). Importantly, the empirical outlier approach does not relay on simulations based on demographic models, an important advantage here as the demographic history of the great apes is yet not fully understood.

5.3.3 Shared targets of balancing selection between species

Previous studies have highlighted examples of long-term balancing selection that result in the presence of trSNPs between different species (Asthana et al. 2005, Cagliani et al. 2010, Cagliani et al. 2012, Ségurél et al. 2012, Leffler et al. 2013). However, being rare, such examples likely provide a limited view on how selective forces are similar across different species. For instance, it is possible that balancing selection acts on the same locus in different species without giving rise to trSNPs, primarily because selection postdates the splits between different species. To investigate this hitherto underexplored phenomenon, we reasoned that if the targets of balancing selection are often shared between species, windows in the lowest NCD2 quantile of a given species (putatively enriched in targets of balancing selection) are more likely than expected to be present in the lowest NCD2 tail of another species.

We explored every possible pairwise comparison (n=36), with within species targets of balancing selection defined at a threshold of p-value < 0.05. For each comparison we calculated the enrichment of sharing relative to that expected by chance for each of the 20 NCD2 0.05 quantiles, with shared defined as the same window being present in the same NCD2 quantile in both species.

We first asked if, in the balancing selection candidate quantile bin (p-value < 0.05), there is a greater than expected enrichment of sharing. Under a scenario of complete independence, the expected proportion of sharing is simply the quantile width, 0.05 (see Methods), with a relative enrichment of 1. However, the assumption of strict independence among NCD2 values between species is violated by their phylogenetic relationships, as shared ancestry increases the correlation between species in their local patterns of polymorphism. For the expected proportion of sharing we decided to use the averaged proportion of sharing observed across all quantiles per pairwise species comparison. To determine if the enrichment of sharing in the candidate quantile bin is significant, we performed a *Bonferroni* corrected binomial exact test based on the counts of intersecting and total windows. For all pairwise comparisons, we observe a significant enrichment of shared windows in the first (5%) quantile (all $P < 2.2 \times 10^{-21}$, binomial exact test), which contains the strongest candidates for balancing selection in each species (Figure 3 and Table S4). We find that this enrichment is strongest for more closely related (split times < 1 Million years) than distantly related species pairs (Figure 3 and Table S4). The relative enrichment of sharing ranges from 3.24 (Western versus Central or Eastern chimpanzees) to 4.04 (Central versus Eastern Chimpanzees) amongst chimpanzees (Figure S1), is 3.4 among the two gorillas (Figure S2), and is 3.24 among the two orangutans (Figure S2). The next most distant relationships are those between bonobos and chimpanzees. Here relative enrichment ranges from 1.67 to 1.70 (Figure 3). For the more distant phylogenetic relationships, enrichment ranges from 1.17 for Western chimpanzees versus Bornean Orangutan and 1.41 for Western Lowland Gorilla, also versus Bornean Orangutan (Figure 3).

These results would suggest that we detect a significant enrichment of shared targets of balancing selection, even amongst the most distantly related clades considered. However it is unclear whether these levels of sharing are expected under neutrality. Though coalescent simulations have inherent limitations, in this case they can give us qualitative insights into how phylogenetic relationships influence the enrichment of sharing under neutrality. In order to address this question, we simulated 2.0 million 3.0 kb neutral windows per Great Ape species, using realistic demographic scenarios for great ape population history (Prado-Martínez et al. 2013). In the simulated data we compared the enrichment of sharing in the first NCD2 bin (p-value < 0.05), as we did in real data. As in real data, enrichment of sharing is highest amongst the species pairs with split times < 1 Million years.

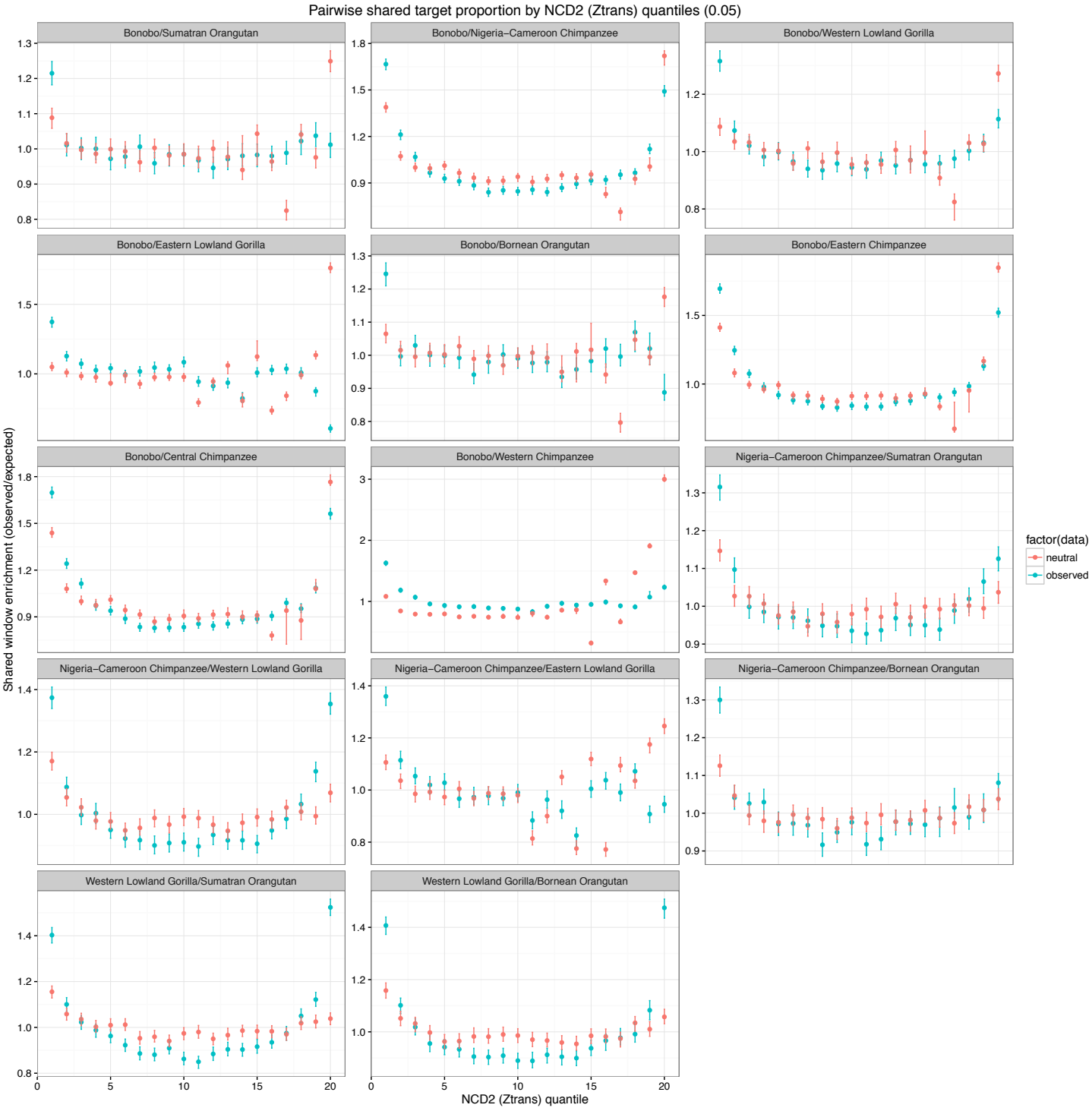
To compare our observed and simulated data, we bootstrapped (n=1000) the sharing enrichment for all quantiles, to estimate 95% confidence intervals of the estimate of the enrichment of sharing, for both observed and simulated data (Methods). In comparisons among chimpanzees, between the gorillas and between the orangutans, we find that the 95% CIs for both observed and simulated candidate quantiles are outside the 95% CI range of the other NCD2 quantiles (Figures S2 and S3). This suggests that much, but perhaps not all, of sharing between closely related species pairs could be explained by neutral evolution. Nonetheless, we only observe higher enrichment under neutrality versus our observed results once, between the two Orangutans (Figure S2). To determine if an enrichment of sharing observed in real data is greater than that observed in the simulations, we propose the following criteria: the 95% CI of the real data candidate quantile is outside the range of all other observed quantiles, while the simulated candidate quantile is within the range of other simulated quantiles. This implies that the observed sharing is not easily explained

under neutrality. On this basis, we find 13 out of the 28 pairwise comparisons showing significant enrichment. Amongst these are all comparisons with Bonobo. These comparisons range split times from greater than 1 Million to up to 10 Million years, and provide evidence for either long term balancing selection, shared targets of balancing selection, or both. Incidentally, bonobos are one of the species where NCD2 has the highest power.

We can further address the extent and timescale of shared balancing selection among species by exploring the extent of sharing as we move up the Great Ape phylogeny. These results are presented in Table 1. At each level of the phylogeny we only consider NCD2 genomic windows analyzed in all species considered at that level, and use p-value < 0.05 as indicative of signatures of balancing selection. We begin within Chimpanzees, as for the current data set this is the only grouping with more than two species with split times of less than one Million years. Of 1,267,249 NCD2 windows, 4,387 are shared by all four chimpanzees, compared to a random expectation of 8 windows assuming completely independence (which shared ancestry obviously violates). This is 3.8 times higher than we observe in neutral simulations (where we only observe 1,549 intersecting windows). Stepping out to consider all of the *Pan* genus (chimpanzees and bonobos), of 1,252,287 NCD2 windows 502 windows are shared. In all African Great Apes 32 windows are shared out of 1,219,989 windows and, in all non-human Great Apes, seven out of 1,198,033. Again, the numbers of shared targets of these three comparisons are both higher than random expectations assuming independence (for each the expectation is less than one window) and higher than we see under neutrality (Table 1). Furthermore only within *Pan* is there more than one window shared under neutrality, as we found 80 intersecting windows in

neutral simulations. For the African Ape and all Great-Ape comparisons, we also studied the sharing at the $p < 0.1$ level. We find 141 and 19 windows shared respectively, compared to only four and zero windows under neutrality.

Unfortunately both phylogenetic distance and geography are nested within the Great Ape phylogeny, thus this analysis cannot reveal whether sharing results from constant long term balancing selection or convergent balancing selection. We note however that of the five genes (*HLA-DRB1*, *HLA-DRB5*, *HLA-DQB1*, *ATN1* and *GARNL4*) associated with the seven windows shared across all great apes at the $p < 0.05$ level, three are HLA genes, a gene family known to harbor haplotypes that are shared across species over long evolutionary time (Klein et al. 1993; Asthana et al. 2005; Loisel et al. 2006; Ségurél et al. 2012; Leffler et al. 2013; Teixeira et al. 2015).


Pairwise shared target proportion by NCD2 (Ztrans) quantiles (0.05)
Bonobo/Sumatran Orangutan
Bonobo/Nigeria–Cameroon Chimpanzee
Bonobo/Western Lowland Gorilla
Bonobo/Eastern Lowland Gorilla
Bonobo/Bornean Orangutan
Bonobo/Eastern Chimpanzee
Bonobo/Central Chimpanzee
Bonobo/Western Chimpanzee
Nigeria–Cameroon Chimpanzee/Sumatran Orangutan
Nigeria–Cameroon Chimpanzee/Western Lowland Gorilla
Nigeria–Cameroon Chimpanzee/Eastern Lowland Gorilla
Nigeria–Cameroon Chimpanzee/Bornean Orangutan
Western Lowland Gorilla/Sumatran Orangutan
Western Lowland Gorilla/Bornean Orangutan
Shared window enrichment (observed/expected)
NCD2 (Ztrans) quantile
factor(data)
neutral
observed

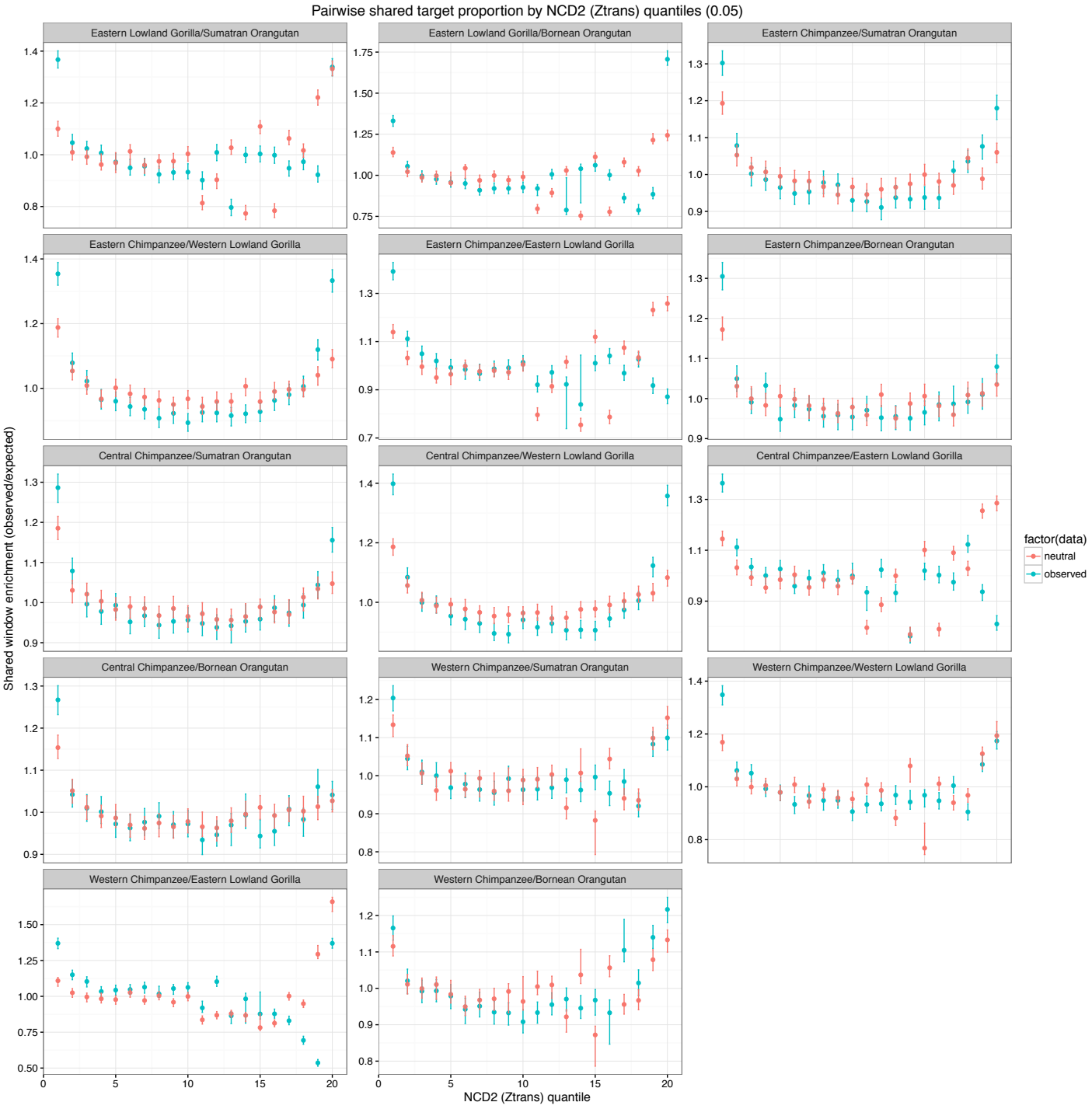


Figure 3: Enrichment of shared windows across NCD2 quantile bins of 0.05 between pairs of great ape species. Comparisons between distantly related species (split time > 1 My) are shown in blue (observed in empirical data) and red (observed in neutral simulations).

Comparison	Observed					Neutral				Obs/Sim Enrichment
P<0.05	Total	Intersect	Expected	Enrichment	p-value	Total	Intersect	Expected	Enrichment	
All Chimpanzees	1,267,249	4,387	8	548	$< 1 \times 10^{-10}$	1,682,821	1,549	11	141	3.8
Chimpanzees/Bonobo	1,252,287	502	<1	502*	$< 1 \times 10^{-10}$	1,657,810	80	<1	80*	665
African Great Apes	1,219,989	32	<1	32*	$< 1 \times 10^{-10}$	1,490,814	1	<1	1*	39
All Non-human Great Apes	1,198,033	7	<1	7*	$< 1 \times 10^{-10}$	1,364,007	0	<1	0*	Inf
P <0.1										
African Great Apes	1,219,989	141	<1	141*	$< 1 \times 10^{-10}$	1,490,814	4	<1	4*	43
All Non-human Great Apes	1,198,033	19	<1	19*	$< 1 \times 10^{-10}$	1,364,007	0	<1	0*	Inf

- Both observed and neutral simulation expected number of windows < 1, but are rounded up to 1 to calculate the enrichment.

Table 1: Enrichment in proportion of shared windows in the first NCD2 quantile (P < 0.05 and P<0.1) among different groupings of great ape species. Comparisons within chimpanzee, *Pan*, African great apes and all great apes species are shown. Total number of windows analyzed, number of intersecting windows in the first NCD2 quantile, expected number of intersections by chance, enrichment of sharing and p-value are illustrated for both empirical data and neutral simulations.

5.3.4 The putative targets of balancing selection

In order to understand how balancing selection affects function in the genome we first estimated whether the top candidate windows (P<0.001) in each species tend to overlap with genic regions of the genome (we take the start and end coordinates of hg18 annotated Ensembl genes), by using the R package GenometriCorr (Favorov et al. 2012) (details in Methods). The results are presented in Table 2.

species	Projection test lower tail	Projection test p-value	Abs. distance lower tail	Abs. distance p-value
Nigeria-Cameroon Chimpanzee	FALSE	9.4e-08	TRUE	<2e-04
Central Chimpanzee	FALSE	2.1e-11	TRUE	<2e-04
Eastern Chimpanzee	FALSE	7.5e-12	TRUE	<2e-04
Western Chimpanzee	FALSE	2.2e-16	TRUE	<2e-04
Bonobo	FALSE	0.0000	TRUE	<2e-04
Western Lowlan Gorilla	FALSE	6.7e-06	TRUE	<2e-04
Eastern Lowlan Gorilla	FALSE	2.9e-09	TRUE	<2e-04
Sumatran Orangutan	FALSE	0.029	TRUE	<2e-04
Bornean Orangutan	FALSE	1.3e-09	TRUE	<2e-04
Chimpanzee	FALSE	0.0000	TRUE	<2e-04
Pan	FALSE	2.2e-12	TRUE	<2e-04
African Apes	FALSE	2.8e-05	TRUE	<2e-04
Great Apes	FALSE	0.0006	TRUE	<2e-04

Table 2: Overlap between candidate windows for balancing selection in great apes and genic regions of the genome. The results for the three different test implemented, as well as their respective p-value, are shown for each species considering the entire genome.

We observe a significant excess of overlap of candidate windows with genes in all species (*projection test,* all $P < 0.029$). Furthermore, candidate targets of balancing selection fall closer to genes than expected by chance (*absolute distances test,* all $P < 2.0 \times 10^{-4}$; Table 2). The observation that candidate windows overlap with genes significantly more often than expected does not provide conclusive evidence on how balancing selection affects the evolution of these genes, since selection can affect regulatory functions rather than targeting protein-coding variation.

In order to clarify the role of balancing selection in shaping the evolution of these genes, we use RegulomeDB (Boyle et al. 2012) scores. Specifically, we compared the proportion of sites showing evidence for regulating gene expression in humans

between candidate windows ($P<0.001$) and the remainder of the genome. In order to obtain a background set for the comparison, we performed 1,000 permutations of non-significant windows in each species. We considered the combined proportion of RegulomeDB scores of 1 and 2 (that have the strongest evidence for regulating gene expression), and 7 (positions in the human genome with no evidence for a regulatory role). We observe a significant depletion of RegulomeDB scores of 1 and 2, as well a significant excess of RegulomeDB scores of 7, among candidates for balancing selection (Figure 3). The results are consistent across all species.

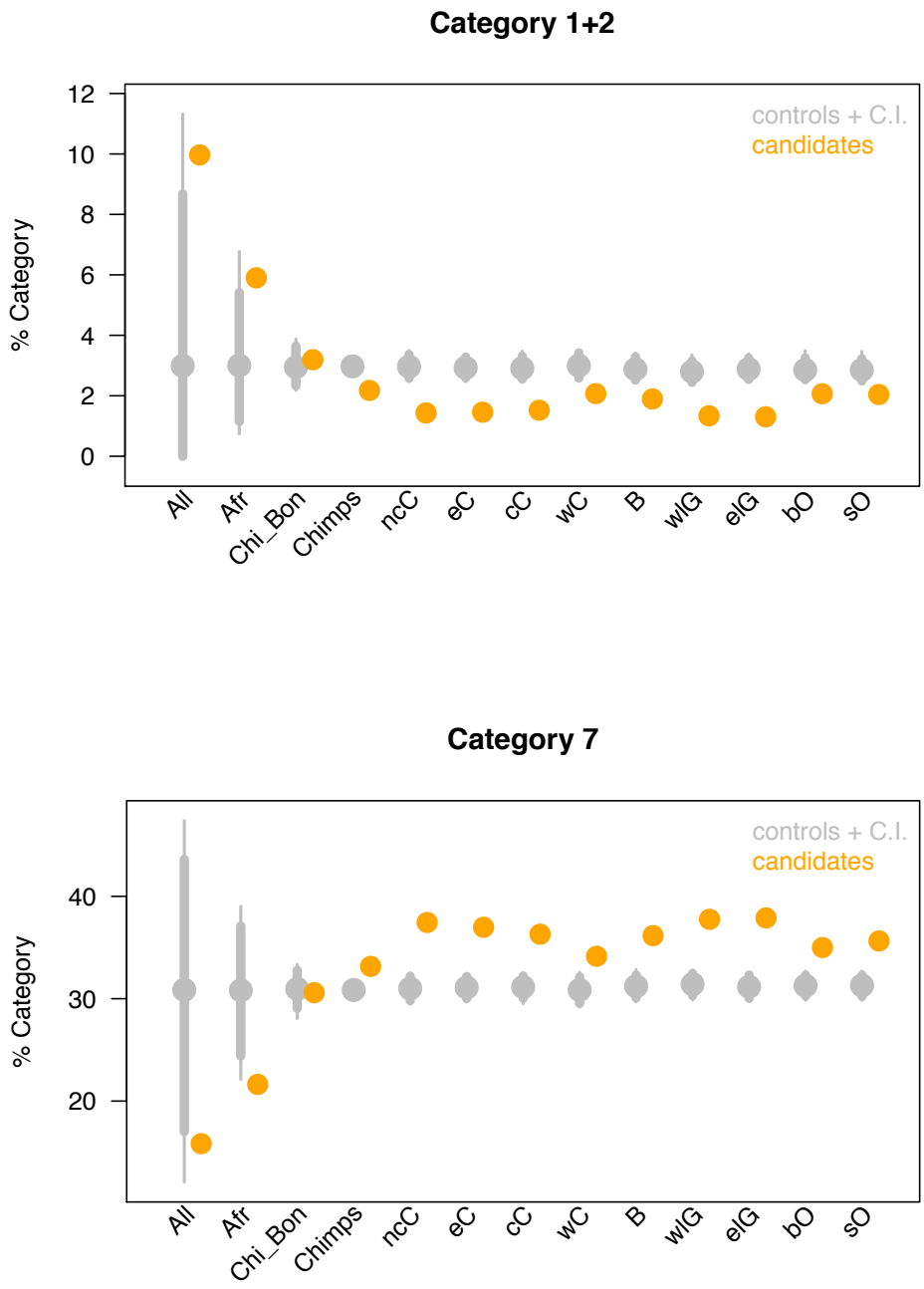


Figure 4: Proportion of RegulomeDD scores of 1+2 and 7 in candidate (yellow) and randomly sampled windows (grey). Different sets of candidate windows in species and groupings are illustrated as follows: All = Great Apes; Afr = African Great Apes; Chi-Bon = Pan clade; Chimps = Chimpanzee; ncC = Nigeria-Cameroon chimpanzee; eC = eastern chimpanzee; cC = central chimpanzee; wC = western chimpanzee; B = bonobo; wlG = western lowland gorilla; elG = eastern lowland gorilla; bO = bornean orangutan; sO = Sumatran orangutan.

Given that we observed a significant proportion of shared targets of balancing selection between different species, which may putatively represent older instances of balancing selection, we replicate the analyzes using candidate windows that are shared in different great apes ($P<0.05$). Specifically, we analyze shared windows in all chimpanzee subspecies, in the *Pan* clade (chimpanzee and bonobo), in African apes (chimpanzee, bonobo and gorilla), and in all great apes (chimpanzee, bonobo, gorilla and orangutan). Once again, we observe a significant overlap between all these sets of candidate windows for long-term balancing selection and genic regions, as indicated by the *projection* (all $P < 2.8$ x 10-5) and *absolute distance* (all $P < 2.0 \times 10^{-4}$) tests (Table 2).

Interestingly, analyses of RegulomeDB scores show that shared targets of long-term balancing selection may be in fact affecting gene regulation. In fact, we observe a continuous increase in the proportion of RegulomeDB scores of 1 and 2 (and a decrease in scores of 7) when restringing to candidate windows shared by more species. Specifically, if we analyze candidate windows shared by all chimpanzees, we obtain similar results as for each species separately, but by simply restringing the analysis to windows that are also candidates for balancing selection in bonobo the proportion of putatively regulatory sites within those windows increases (Figure 4). Moreover, this pattern is even stronger for candidate windows in all African apes in windows shared by all great ape species, where candidate targets of balancing selection fall in the top 5% of the background distribution in the proportion of RegulomeDB categories 1 and 2 sites. These results seem to indicate that unlike species-specific targets of balancing selection, shared instances of balancing selection may affect regulation of gene expression.

5.3.5 Gene Ontology analysis

In order to assess whether biological functions and pathways are targeted by balancing selection, we use GOWINDA (Kofler and Schlötterer 2012) to perform Gene Ontology (GO) analysis on species-specific candidate windows ($P<0.005$), as well as in candidate windows shared by different species ($P<0.05$), using the groupings as indicated above. Unequivocally, the results are dominated by the presence of HLA genes. In virtually all species, the top GO categories indicate immune response, particularly antigen processing and presentation, which is the main function of the MHC complex. Interestingly however, there is one exception. In western chimpanzee, despite the abundant presence of MHC-related GO categories, the most significant GO category is *hemoglobin complex*. A total of 5 hemoglobin genes appear as candidate targets of balancing selection in western chimpanzee, *HBA1* and *HBQ1* in chromosome 16, and *HBD*, *HBE1*, and *HBG2* in chromosome 11.

5.3.6 Targets of balancing selection in all great apes

We further investigated candidate windows that show evidence of balancing selection (5% tail) in all great ape species, and uncovered a total of 7 windows overlapping five different genes: *HLA-DRB1*, *HLA-DRB5*, *HLA-DQB1*, *ATN1* and *GARNL4*. Interestingly, all windows overlap regions of the genome with evidence for regulation of gene expression. While these windows are likely to represent outstanding targets of balancing selection, one possible confounder is that they are, instead, duplicated regions in great apes, or deletions in human. Even though we implemented stringent filtering criteria that should ensure the exclusion duplicated regions of the genome, we nevertheless explored whether differences in coverage depth exist between these

windows and the remainder of the genome. We show, for all individuals across all species that these windows do not include coverage outliers and show very similar coverage patterns as the rest of the genome (Figure S3). This observation is reassuring and makes it much more likely that these represent true candidate targets of balancing selection.

Of the seven windows, one encompasses the first exon of *HLA-DRB1* and *HLA-DRB5*, and is centered just upstream both genes (chr6:32557270-32560270). The region shows an abundance of transcription factor (TF) binding associated with histone modifications and, therefore, likely includes the promoter region of both genes. Additionally, we found another window 10kb upstream that also shows evidence of histone modification and includes the binding region of the TF YY1. A third and forth windows (chr6:32567770-32570770 and chr6:32633770-32636770) are adjacent to each other and are located just upstream *HLA-DQB1*. The region also includes abundant TF binding and evidences histone modification, whereby is also very likely to include the promoter region of the gene. Another two windows are overlapping with each other (chr12:7044503-7047503 and chr12:7046003-7049003) and, apart from including the exons 5, 6 and 7 of the gene *ATN1*, have strong evidence for histone modification and also include abundant binding of various TFs. Finally, another window lies just upstream the gene *GARNL4/RAP1GAP2* (chr17:2698703-2701703) and includes the 5' UTR of the gene, also showing a high degree of TF binding.

Additionally, we observe signatures of balancing selection in the gene *OAS1* in all chimpanzee subspecies but western ($P<0.001$). Interestingly, this gene has been previously described as a target of balancing selection in both central and eastern chimpanzees (Ferguson et al. 2012). Our evidence suggest that balancing selection

also affects the evolution of this gene in Nigeria-Cameroon. Furthermore, *OAS1* is among the genes containing one trSNP in humans, chimpanzees and bonobos (Teixeira et al. 2015), although segregating at very low frequencies in humans and bonobos.

5.4 Discussion

This work provides the first genome-wide scan for targets of long-term balancing selection maintaining intermediate-frequency alleles in all of the major non-human great ape clades: *Pan*, *Gorilla*, and *Pongo*. Our strategy relies on using the NCD2 statistic (Bitarello et al. (in prep) to uncover targets of balancing selection. NCD2 is designed to be sensitive to the accumulation of polymorphisms segregating at intermediate frequencies, and on the decrease in the number of fixed nucleotide changes to human, both of which are features expected in genomic regions targeted by long term balancing selection.

Our power to detect balancing selection using NCD2 is greatly influenced by two factors: the action of recombination and long-term effective population size (Ne). We addressed both of these through neutral and selection simulations based on recent reports of great ape population history (Prado-Martínez et al. 2013).

We note that the action of recombination through time disrupts linkage between the putatively selected sites and nearby neutral variation (Charlesworth et al. 2006; Leffler et al. 2013). Therefore, the length of the region carrying signatures of balancing selection should decrease as a function of the time since the onset of selection. Our power analysis suggested that using a 3kb sliding window approach maximized our ability to detect signatures of long term balancing selection.

We also demonstrate that NCD2 performs better in species where Ne is low, which comes as no surprise given that populations with low Ne have an overall depletion of genetic diversity in the genome. Hence, if balancing selection is effective in maintaining advantageous diversity in species with low Ne, these regions appear more easily as outliers in a genome-wide scan, as the TMRCA of a region under

balancing selection is likely much older than the average TMRCA of the genome in species with low Ne. Overall, we show that NCD2 has high power to detect instances of long-term balancing selection in all great ape species that is at least as old as 3.5 million years.

It may be possible to detect shared targets of balancing selection in different species because balancing selection can maintain advantageous polymorphisms in populations for millions of years (Charlesworth 2006, Andrés 2011, Key et al. 2014). The timescale of selection impacts the levels of genetic diversity in a region and results in different signatures in the genome. For example, if balancing selection is old enough and predates the split of different species, it can lead to the presence of trSNPs at the selected loci. In fact, previous studies attempted to uncover shared targets of balancing selection by uncovering instances where long-term balancing selection maintains trans-species polymorphisms between humans and other great apes (Klein et al. 1993; Asthana et al. 2005; Cagliani et al. 2010; Cagliani et al. 2012; Ségurél et al. 2012; Leffler et al. 2013; Teixeira et al. 2015). Contrarily, targets of balancing selection might be shared across species due to convergent evolution, in which case the onset of selection postdates the split between the species. We provide evidence that targets of balancing selection are shared across different great ape species beyond the expectations of shared population history, particularly for species that have diverged millions of years in the past. Specifically, we see a significant enrichment on the proportion of shared windows in the tail of the empirical distribution ($P<0.05$), which includes the strongest candidates for balancing selection. Albeit we can replicate this observation using neutral simulations of great ape population history in species belonging to the same clade, we cannot reproduce it for

more distantly related species, namely for comparisons between chimpanzees/bonobo with gorillas/orangutans and between gorillas and orangutans.

While it is possible that technical artifacts may contribute to the observed unexpectedly high proportion of shared targets of balancing selection in different species, this possibility seems highly unlikely. The major possible factor could be the presence of duplications in the great apes that are not shared with humans, or analogously, human specific deletions that could affect mapping and thus create artifact SNPs. Annotated duplications are excluded from the analysis by removing genomic segments annotated as segmental duplications in the hg18 reference genome as well as by removing copy-number variants in all species that were annotated from this very same dataset (Sudmant et al. 2013). Moreover, even if some un-annotated CNVs persisted after these filters, we put in place strict coverage and heterozygous filters, designed to remove regions of the genome where paralog sequences may map to the same hg18 region. Additionally, we remove positions where all individuals are heterozygous. In fact, we show that among the strongest candidate windows in our scan (those showing evidence of balancing selection in all species) coverage depth is within the range of the remainder of the genome. Together, these observations make it extremely unlikely that technical artifacts are a major problem in our analysis.

Interestingly, we observe that targets of balancing selection that are shared across distantly related species have significant excess of sites with putatively regulatory function. A regulatory role for long-term balancing selection was proposed by Leffler et al. based on patterns of shared polymorphism between human and chimpanzee, and our results are in agreement with such observations (Leffler et al. 2013). Perhaps more intriguing is the fact that species-specific candidates, where balancing selection is perhaps more recent, are targeting genic regions of the genome without evidence

for an influence in gene expression. This observation is quite remarkable in the sense that these regions are expected to evolve under strong selective constraint and, therefore, a scenario where relaxed constraint is responsible for the signatures observed seem quite unlikely. It is instead much more plausible that these results provide evidence for an unexpectedly high proportion of loci evolving under balancing selection in species that diverged millions of years ago, and demonstrate that targets of balancing selection are shared between species. It is possible to envision a scenario where environmental pressures are constant for millions of years and shape the evolution of gene expression similarly in different species, likely since before their split.

Our GO analysis revealed that the vast majority of categories associated with candidate targets of balancing selection in great apes are located in the MHC cluster in chromosome 6. If indeed balancing selection mostly affects the antigen presentation pathway, then it is possible to speculate that it does so by maintaining advantageous polymorphisms in complementary strategies: at the protein-coding level as a direct response (more recent, species-specific balancing selection), and at the regulatory level turning genes off and on (older, shared balancing selection across species).

In fact, we uncover seven outlier windows that are shared across all great apes species. These results are as unexpected as they are remarkable in illustrating examples where selective pressures are the same across species, and it is even possible that the onset of selection occurred at the common ancestor of these species. Again, coverage data shows no indication that these may represent copy number variants in the genome, although detailed experimental evidence is necessary to fully discard that possibility. While the focus of this study are the general patterns of

balancing selection, these loci are obviously of great potential interest as striking candidate targets of balancing selection. Interestingly, two of these windows are located just upstream *HLA-DRB1* and *HLA-DRB5*, in regions that show strong evidence of gene expression regulation. Moreover, one of these windows overlaps with the binding site YY1 (10kb upstream *HLA-DRB1*), a TF that is the main known repressor of HIV-I transcription (Margolis et al. 1994; Pereira et al. 2000; Coull et al. 2000; Bernhard et al. 2013). Recently, alleles associated with differences in gene expression of *HLA-DRB1* were found to be associated with differences in susceptibility to infection by HIV-I in humans (Ranasinghe et al. 2013). While this example is particularly interesting, the complex nature of antigen processing and presentation leaves the question open as to why these particular regions of the genome show signatures of balancing selection. Apart from *HLA-DRB1* and *HLA-DRB5*, we found an additional two (consecutive) windows that are located just upstream *HLA-DQB1*, in a region with evidence for TF binding and histone modification, also suggestive that polymorphisms in the region might be associated with differences in gene expression, particularly considering that the region is most likely overlapping with the promoter of the gene.

Balancing selection has previously been shown to affect the evolution of MHC genes and is in fact responsible for maintaining trSNPs in different species (Graser et al. 1996; Kikkawa et al. 2009; Sutton et al. 2013; Cutrera et al. 2007; Klein et al. 1993; Asthana et al. 2005; Loisel et al. 2006; Ségurél et al. 2012; Leffler et al. 2013). Even though we cannot test this explicitly using our approach, combined evidence with other studies suggests that it is perfectly possible that significant windows in these HLA genes represent cases where balancing selection has been maintained since the common ancestor of all great apes.

Moreover, we found signatures of balancing selection in two overlapping windows in the gene *ATN1* (atrophin-1), which apart from including three different exons of the gene are also overlapping intronic regions with evidence for histone modification. This candidate region overlaps a known trinucleotide (CAG) repeat that causes dentatorubral-pallidoluysian atrophy (DRPLA) in humans, a neurodegenerative disorder which clinical manifestations that include epilepsy, ataxia and dementia (Nagafuchi et al. 1994, Koide et al. 1994). Finally, we also uncovered a window showing signatures of balancing selection in all great apes that locate right upstream the gene *GARNL4/RAP1GAP2*, and that includes the first exon of the gene. This region once again shows evidence for TF biniding and histone modification and its variants are likely to be important in regulating the expression of *GARNL4/RAP1GAP2*. Interestingly, this gene regulates secretion of dense granules of platelets at sites of endothelial damage. A growing body of evidence suggests an important role for platelets in immune response (von Hundelshausen and Weber 2007, Semple et al. 2011), as they can directly interact with viruses (Assinger 2014). In fact, while viruses can induce thrombocytopenia (a drop in platelet count), platelets can activate immune cells and show antiviral and antimicrobial activity (Assinger 2014, Yeaman 2014). Whether balancing selection acts on a gene involved in platelet formation due to the immune-like nature of platelets remains, nevertheless, a matter of speculation and further work is necessary to ascertain whether such link does exist. Finally, we uncovered signatures of balancing selection in the antiviral gene *OAS1* in central, eastern and Nigeria-Cameroon chimpanzees. Interestingly, another study using independent sampling and a different approach has demonstrated that *OAS1* is a target of balancing selection in central and eastern chimpanzees (Ferguson et al. 2012), which shows the power of NCD2 to uncover true targets of balancing selection

in the genome of great apes. Interestingly, Teixeira et al. (2015) found a trSNP segregating in *OAS1* in humans, chimpanzees and bonobos, although only segregating at intermediate frequency in chimpanzees.

This works presents a first attempt focused at understanding how balancing selection has shaped the evolution of great apes. We found evidence that a high proportion of targets of balancing selection is shared by different species and mostly affects genic regions of the genome. Furthermore, we provide evidence that regions evolving under balancing selection in different species seem to have a role in regulating gene expression. Five different genes emerge, from this study, as interesting candidates of long-term balancing selection in all non-human great apes, whose signatures detailed studies will help better understand.

5.5 Materials and Methods

5.5.1 Samples

We analyzed a total of 73 samples from 9 different subspecies of great apes from the Great Ape Genome Project (Prado-Martínez et al. 2013). The samples analyzed in our study include 4 central chimpanzees (*Pan troglodytes troglodytes*), 4 western chimpanzees (*P. t. verus*), 6 eastern chimpanzees (*P. t. schweinfurthii*), 10 nigeria-cameroon chimpanzees (*P. t. ellioti*), 13 bonobos (*Pan paniscus*), 3 eastern lowland gorillas (*Gorilla beringei graueri*), 23 western lowland gorillas (*Gorilla gorilla gorilla*), 5 bornean orangutans (*Pongo pygmaeus*) and 5 sumatran orangutans (*Pongo abelii*). These samples virtually comprise the totality of available genomes from Prado-Martínez et al. (2103). Following that study, we excluded four related western lowland gorillas (Bulera, Kowali, Suzie and Oko), and one western chimpanzee (Donald). Furthermore, we do not consider cross-river gorilla in this study since we lack population data (only one individual was sampled by Prado-Martínez et al. (2013)).

We downloaded VCFs files for each subspecies from https://eichlerlab.gs.washington.edu/greatape/data/VCFs/. Whole genome sequencing was performed using Illumina HiSeq 2000, read-mapping to hg18 was done with BWA (Li and Durbin 2009), and the Genome Analysis ToolKit (GATK) (McKenna et al. 2010) was used for variant calling, as described in Prado-Martínez et al. (2013).

5.5.2 Data filtering

The VCFs files were filtered according to several quality-control criteria, following Prado-Martínez et al. 2013. SNPs were removed if they had

- DP < (mean_read_depth/8.0) || DP > (mean_read_depth*3)
- QUAL < 33
- FS > 26.0
- Sites within 5 bp of a reported indel
- MQ < 25
- MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1

Furthermore, Prado-Martínez et al. 2013 predicted segmental duplications based on read-depth counts using mrsFAST and excluded variants overlapping with those regions. Finally, due to problems with contamination, Prado-Martínez et al. (2013) implemented an allele balance filter that excluded 10% of heterozygous positions with the highest allele imbalance in each species.

We implemented several additional filters on the data to remove false SNPs as a consequence of unannotated duplicated regions. We started by excluding SNPs overlapping known segmental duplications and repeat-regions in the hg18 genome (downloaded from https://genome-euro.ucsc.edu/), and annotated copy-number variants in any of the 9 subspecies (Sudmant et al. 2013). Additionally, we excluded positions that are coverage outliers in any species. To achieve that, we used the coverage distribution per position of each individual, and flagged SNPs with less than 5X coverage or that lie in the upper 10% of the coverage distribution of that individual. We then excluded all positions flagged in more than one third of the individuals in each species. Finally, we excluded all SNPs for which genotype

information is missing in more than one third of the individuals in each species, as well as those for which all genotyped individuals are heterozygous. Specifically, we used a custom *perl* script to count, separately for each species, the number of heterozygous individuals per polymorphic position. We noticed even after filtering, that sites for which every individual had a heterozygous genotype were often clustered together. Sites with a 100 per cent heterozygosity rate were therefore excluded.

5.5.3 Uncovering targets of balancing selection using NCD

To detect targets of long-term balancing selection in the genome of great apes we used a statistics presented in Bitarello et al. (*in prep*) that measures the degree to which the local site-frequency spectrum of a particular locus deviates from a specified target allele frequency, defined by the authors as a "Non-Central Deviation" (Bitarello et al. *in prep*). Under balancing selection, this target frequency can be regarded as the frequency at which the action of selection maintains selected alleles in the population. Bitarello et al. propose two different implementations of the statistic, NCD1 and NCD2. Here we use the NCD2 implementation, which relies on information provided by the frequency of polymorphic sites (p_i) and the number of fixed differences (FDs) in a given locus,

$$NCD2_{tf} = \sqrt{\frac{n_{fd} \cdot (0 - tf)^2 + \sum_{i=1}^{n} (p_i - tf)^2}{n_{fd} + n}}$$ (from Bitarello et al. *in prep*)

where $i = \{1,2,3,...n\}$ is the *i*-th polymorphism in the locus, p_i is the minor allele frequency (MAF) for the *i-th* polymorphism, *tf* is the target allele frequency with respect to which the deviations of the observed alleles frequencies are computed (Bitarello et al. *in prep*), and n_{fd} is the number of FDs in the locus.

We used a target allele frequency of 0.5, and considered the folded SFS (given that frequency at bi-allelic sites is complementary) for NCD calculations, whereby our analysis bounds NCD $_{value}$ between 0 and 0.5. NCD was calculated in 3kb sliding (1.5kb) windows of the genome, as in Bitarello et al., (in preparation) 3kb maximizes the power of the statistic. For simplicity we considered only bi-allelic sites.

Bitarello et al., (in preparation) showed that the NCD statistics have equal or higher power than existing methods to identify signatures of balancing selection such as Tajima's D (Tajima 1989), HKA (Hudson et al. 1987), these two tests combined, or T1 and T2 (DeGiorgio et al. 2014).

5.5.4 NCD variance and the number of *IS* per window

Noise in the estimate of the NCD statistic per window is not independent from the number of informative sites (SNPs plus FDs). As expected, as a group the windows with low number of *IS* have the highest variance in the NCD statistics (Figure 2 and Table S2). We account for this dependency in two ways: with an additional filter, and with a method to assign p-values that accounts for the number of *IS* in the window.

To the exclude windows showing the lowest number of *IS* and he highest NCD variance, we removed windows that have a number of *IS* that falls in the 5% quantile of the distribution for each species. 5% was chosen because we observe the biggest drop in NCD2 variance after exploring several possible values (5%, 10%, 15%, 20, 25% and 30%, Figure 2 and Table S2).

5.5.5 Defining candidate windows

To minimize the impact that the number of *IS* has in the NCD2 p-value, we normalized the NCD2 $_{value}$ of each window according to its number of *IS*. First, we compute the mean NCD2 $_{value}$ of all windows with a certain number of *IS*. Then, for each window we compute a standardized distance measure between its observed NCD2 $_{value}$ and the average NCD of all windows with the same number of *IS*. This distance (Z_{ft}) is given by:

$$Z_{tf} = \frac{NCD_{ft} - \overline{NCD}_{bin}}{sd_{bin}}$$ (from Bitarello et al. *in prep*)

We then defined candidate windows by using an empirical p-value approach on the Z_{ft} NCD2 calculations. We considered two separate sets of candidate windows based on two different empirical cutoffs employed: P<0.1% and P<0.5%.

5.5.6 Simulations and power analysis

In order to assess the power of NCD2 to unveil true targets of long-term balancing selection in the genomes of great apes, we performed coalescent simulations under neutrality and selection using the software *msms* (Ewing and Hermisson 2010). We defined a demographic model for great ape population history following Prado-Martínez et al. 2013 PSMC curves, and used a generation time of 20 years, a mutation rate of 10^{-9} mutations per nucleotide site per year, and a recombination rate of 10^{-9} (the values used to perform the PSMC analyses). Since *msms* has problems when simulating balancing selection on several demes, we run pairwise simulations

of each great ape population with human. We performed simulations of different sequence lengths in order to test the power of NCD1 and NCD2 – 3kb, 6kb and 15kb. We used a model of overdominance for simulations with balancing selection and required that the balanced polymorphism was located in the center of the simulated sequence, with the onset of selection occurring 3.5Mya and selection on heterozygote advantage, where the fitness of heterozygous individuals is 100x higher than both homozygous.

The power of NCD to detect target of balancing selection was measured by the relationship between the true and false positive rates (simulations under balancing selection and under neutrality) and represented by receiver operating characteristic (ROC) curves.

5.5.7 Shared targets across species

We calculated the proportion of sharing for each pairwise species comparison using a custom R script. For each comparison, only windows for which a valid Zscore could be computed in both species were considered. For each species, a p-value for each window was calculated using the empirical cumulative distribution function, then windows were binned in 0.05 quantiles of the p-value distribution. Next for each quantile bin, we counted the number of windows for each species (nP1 and nP2) and the number of intersecting windows (nIntersect), defined as windows with the same start and end coordinates (i.e. partially overlapping sliding windows were not counted as shared). The proportion of sharing in the quantile bin is then: nIntersect/mean(nP1, nP2).

The expected proportion sharing for independent data is just the quantile width, 0.05 as follows: The probability that a window is in a particular quantile, q, in both species

is p^2, 0.05 x 0.05 = 0.0025, but the proportion of windows in *q* is, by definition, 0.05. Thus the expectation of sharing is 0.05 x 0.05 / 0.05 = 0.05. However, we empirically calculate the expectation of sharing as the mean of the observed proportion of sharing across all quantile bins, mShare. We calculate a binomial p-value for *greater* than expected sharing at bonferroni corrected alpha 0.05, using the R function binom.test with: x= nIntersect, n=max(nP1, nP2), p= mShare, alternative="greater"), and matching upper and lower confidence intervals using the "agresti-coull" method in the binom.confint R function.

To calculate 95% bootstrap confidence intervals, we sample with replacement from the pairwise matrix of matching NCD windows (i.e. the matching window from species 1 and species 2 are sampled together) with the sample size equal to the observed number of windows, for a total of 1000 bootstrap replicates. The entire proportion of sharing procedure outlined above was then calculated for each replicate, with the 95% confidence interval then calculated from the 0.025 and 0.975 quantiles of the bootstrap distribution.

5.5.8 Proportion of shared targets in neutral simulations

We simulated 3 kb long loci following the neutral joint demographic history of the nine great Apes and a single human chromosome in MSMS on a server cluster for 40 x 50,000 = 2 Million replicates. MSMS output was read into R, and NCD scores per window per species were calculated, using custom R scripts. Before performing the procedure outlined above for analyzing the pairwise proportion of sharing, windows in the bottom 5% quantile for informative sites were removed per species.

5.5.9 Intersects sets of NCD2 candidates

For each intersect set, we first required that each window had a valid Z_{tf} NCD score for each species in the comparison. We then used the p-values calculated from the empirical distribution to find the windows that were below the required threshold in each species of the comparison. For finding union sets, we again only used windows that had a valid Z_{tf} NCD score for each species in the comparison, but took windows with a p-value below the required threshold in a least one of the species.

5.5.10 Gene and Phenotype Ontology

To test whether particular biological processes or functions are preferential targets of the action of balancing selection, we started by defining *candidate* genes as those overlapping a candidate window, regardless of the genic element of overlap.

GO (gene ontology) enrichment analyses were performed using the software GOWINDA (Kofler and Schlöterer 2012), which is design to circumvent several common biases in ontology enrichment analysis, namely gene length (longer genes with more windows have by chance a higher probability of containing a candidate window) and clustering (given that some gene families are located in clusters along the genome).

We manually constructed input files to use with GOWINDA, namely the gene annotation (.gtf) and gene set (list of GO term accessions and associated genes) files. We first downloaded the human genome (hg19) chromosome, start and end coordinates, ensembl gene identifiers and external gene names for each autosomal, known, protein coding gene from ENSEMBL biomart (18,564 genes total; accessed Feburary, 2016). The external gene name is most often used to associate genes with GO terms in databases such as those provided by the Gene Ontology Consortium,

however a fraction of genes in ENSEMBL have more than one ensembl gene ID associated with an external gene name (18,512 unique and 26 duplicated genes). This can result in the same gene being associated with the same GO term more than once, and/or different versions of the gene being associated with different GO terms. This can have an adverse effect on the sampling estimates of GO enrichment, particularly if these "duplicates" are overlapping in the genome.

To resolve these inconsistencies, we first checked whether loci for genes with duplicate names were on the same or different chromosomes. One gene, CKS1B, has an ensembl ID for loci on chromosome 1 and 5. These two versions were renamed CKS1B_1 and CKS1B_5. For the remaining 25 duplicates, we checked if the gene coordinates were intersecting or disjoint. If they were intersecting we merged all gene coordinates to create a single gene entry, taking the minimum start and maximum end positions as the new gene start and end positions, respectively. Disjoint duplicates were simply renamed geneX_a, geneX_b, geneX_c etc. as required. The gtf file was then compiled from this modified gene set.

Also from ENSEMBL biomart, we downloaded GO term accessions and ensembl IDs for each autosomal, known, protein coding gene. We replaced ensembl IDs for gene names, using the modified list from above, removing duplicates so a gene was only associated with a GO term accession once. A custom *bash* script was then used to modify this list into a GOWINDA gene set file.

We ran GOWINDA in *mode:gene* and computed 100,000 simulations for false discovery rate (FDR) estimation. We considered a given category to be significant when FDR<0.05. Since GOWINDA was designed to perform SNP-based analysis, we considered the middle position of every scanned window as the target site (i.e. "SNP"). Because this strategy prevents a gene to be considered as *candidate* if the

middle position of the window does not overlap the gene (even if the window does), we extended gene coordinates by 1,500bp up/down-stream by using the option *updownstream1500*, and therefore considered the correct coordinates of each window. We downloaded the annotation (.gtf) and gene set files needed to run the Gene Ontology from Ensembl (http://www.ensembl.org/), and performed separate analysis for each lineage's sets of candidates at 0.1% and 0.5% windows.

5.5.11 RegulomeDB analysis

We explored putative regulatory functions among targets of balancing selection by using RegulomeDB, which is a SNP-based annotation for known and predicted regulatory elements (Boyle et al. 2012). Because these annotations are predicted based on the human genome, we considered RegulomeDB scores of 1 and 2 together, since these categories are the same except that 1 includes eQTL sites. Together these categories include positions with evidence of eQTL (1), TF binding (1+2), matched TF motif (1+2), matched DNase Footprint (1+2), and DNase peak (1+2), and therefore show the highest potential for involvement in regulatory functions in the genome. By joining both categories only the sequence information is used (not the evidence for eQTL in humans), making this appropriate for use in non-human species. We also considered RegulomeDB category 7, which includes sites with no regulatory annotation.

For all analyzed 3kb windows in each species, we sum the number of SNPs with a given score (1/2 or 7). We then compute the proportion of SNPs with that RegulomeDB score (out of all SNPs) across candidate windows. We compare this value with the expectation based on the analysis of the entire genome, which is given by 1,000 samplings without replacement of the number of candidate windows in each

species from the background set (i.e. all analyzed windows). This provides an empirical p-value on the enrichment (or depletion) of RegulomeDB scores in candidate windows, controlling for the size of the candidate windows set in each species (we considered significance for $p<0.05$).

5.5.12 Enrichment in genic regions

In order to define SNPs as genic or intergenic we used annotations from the 1000 Genomes Project (Abecasis et al. 2012). To test for enrichment in genic regions among targets of long-term balancing selection, we used Genometric Correlation (Favorov et al. 2012), which is available as an R package (http://genometricorr.sourceforge.net/). This software tests whether different genetic features are spatially arranged and correlated within chromosomes, and whether they are independent of each other in terms of their genomic coordinates (for example, one can think of the correlation between genic regions of the genome and GC content). In our case, we aimed at understanding if targets of balancing selection in the genomes of great apes tend to be closer to genes than expected by chance.

The Genometric Correlation software is able to compute several statistics in order to get correlations between the sets in the comparison. It requires the user to first define the *query* set (set of genomic intervals to be tested, in our case candidate windows of long-term balancing selection) and the *reference* set (the features of the genome to which the *query* set is to be compared to). Here, we used the 1) the *absolute distance* test, which compares the minimum absolute distance between *query* and *reference* sets, and uses permutations to obtain a two-tailed p-value that reflects significant proximity or distance between query and reference; and 2) the *projection* test, which compares the distribution of *query* midpoints that overlap *reference* intervals to a null

distribution given by a binomial distribution with probability of success p (where p = coverage of the reference / chromosome length.

We started by defining the genomic position (chromosome, start, end) of every candidate window as *query*, and the genomic position of known human genes (hg18) as *reference*. A complete list of known human genes was downloaded from Ensembl (http://www.ensembl.org/). We then defined a 'pseudo'-hg18 genome based on the filters applied on the SNP data, which includes only regions of the genome where SNP calls were included, and bounded it to the start and end coordinates of the first and last genomic window for which NCD could be calculated in each species. This is a crucial and conservative step in the analysis as it limits the genomic coordinates to where candidate windows can be permuted in each test, and resulted in a smaller in an effective smaller version of the hg18. In order to calculate significance for each of the tests for each species, we ran 5,000 permutations.

5.6 Supplementary Information

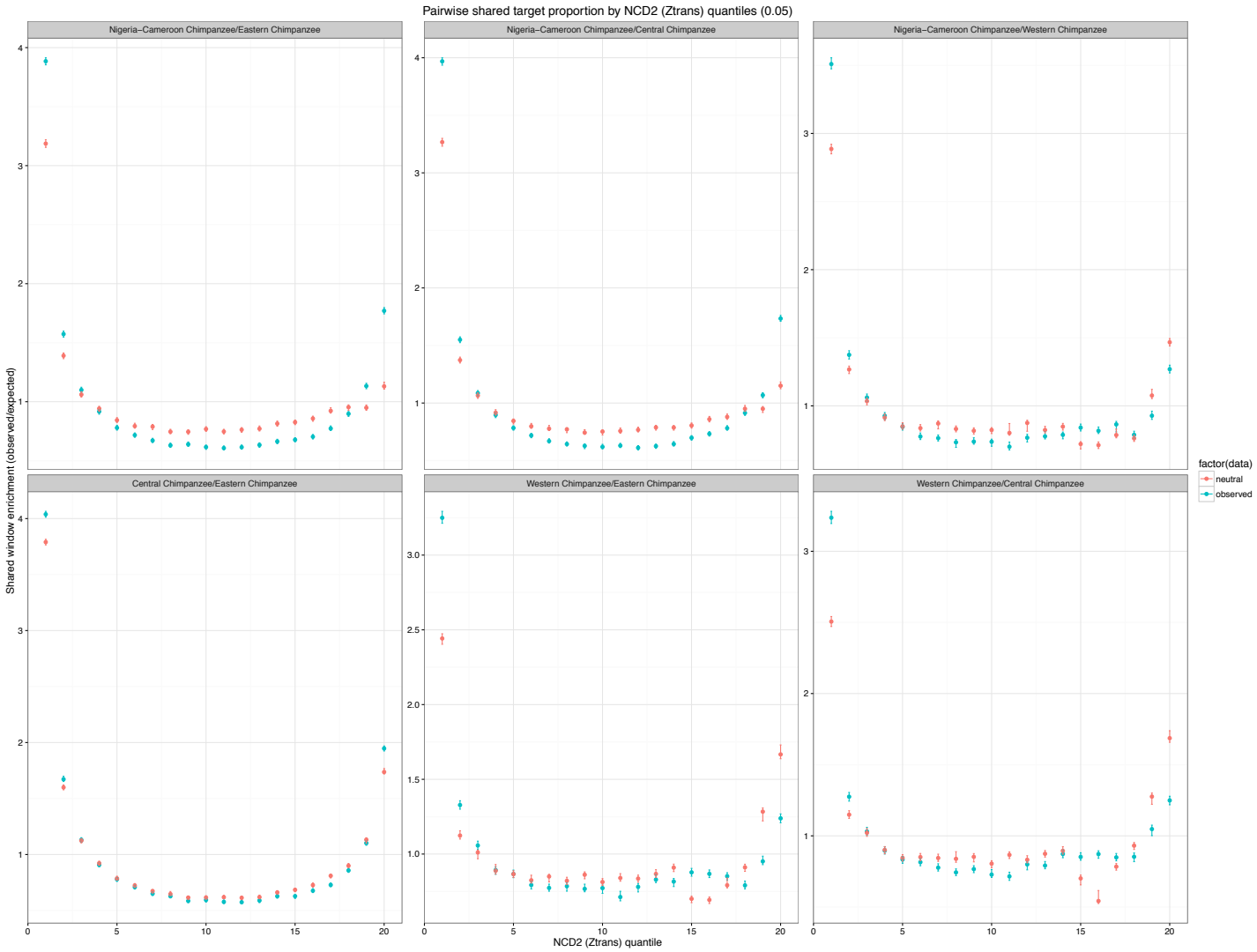


Figure S1: Enrichment of shared windows across NCD2 quantile bins of 0.05 between chimpanzees. Observations based on empirical data are shown in blue, whereas results obtained in neutral simulations are shown in red.

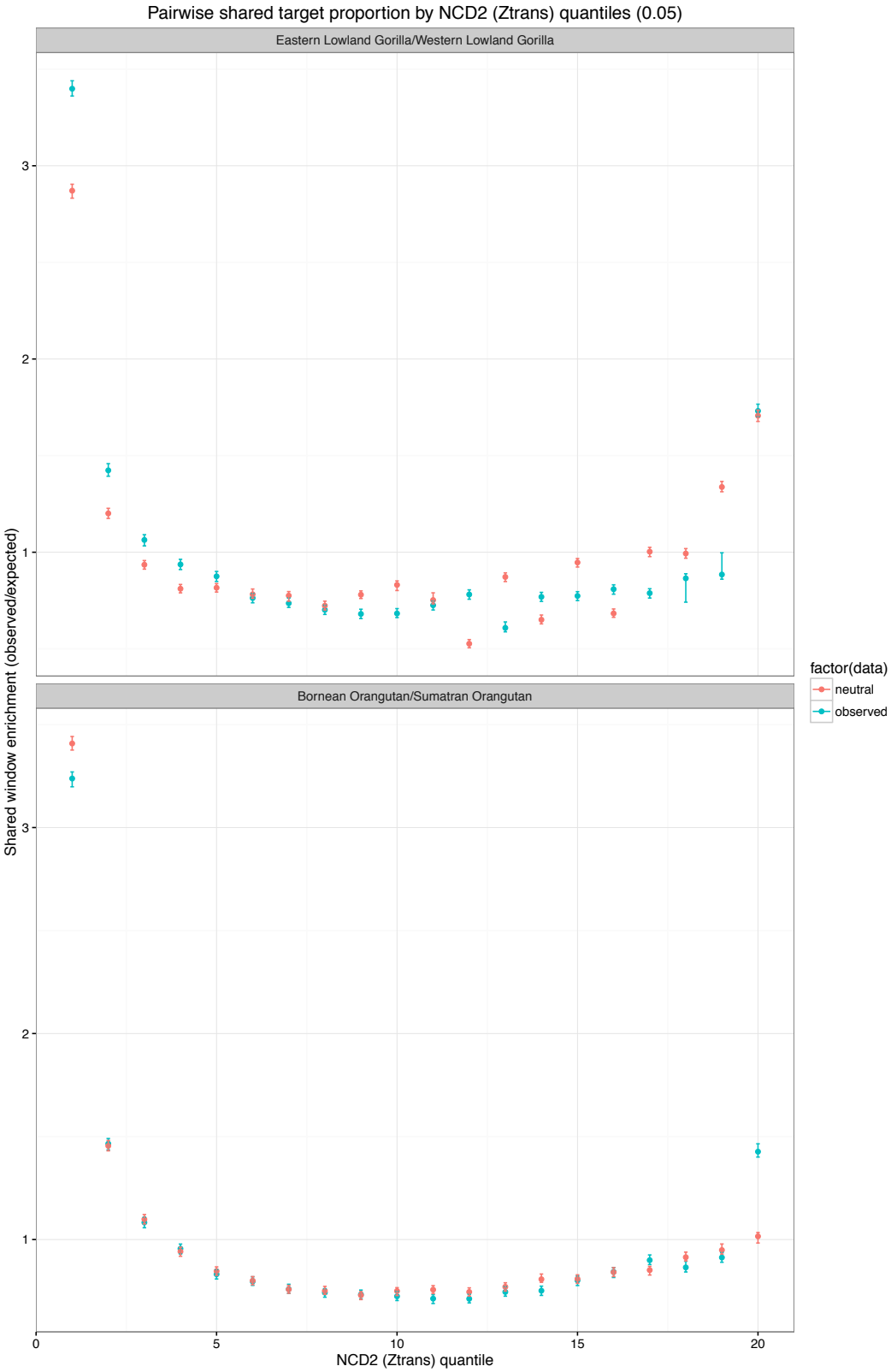
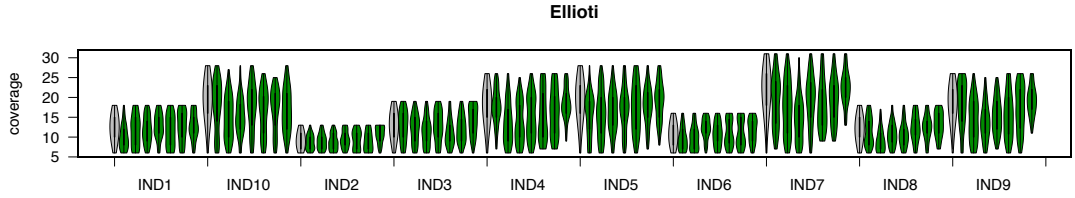
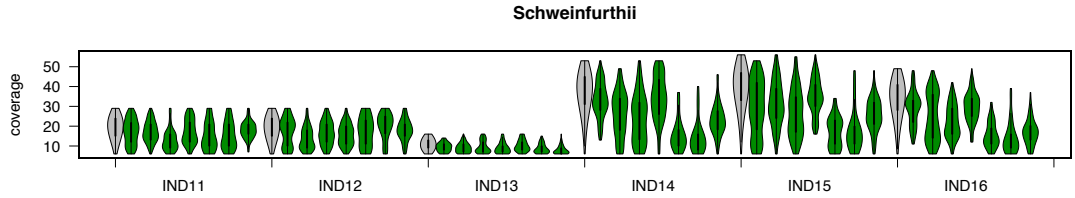
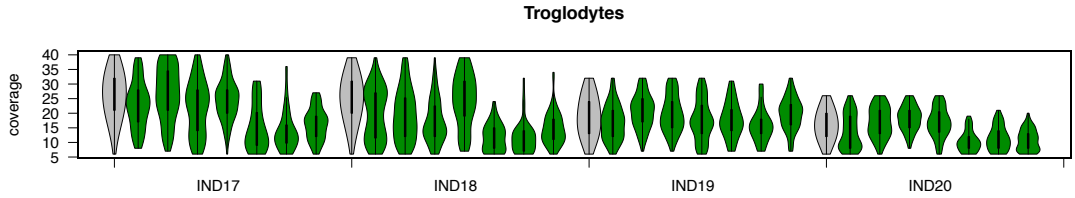
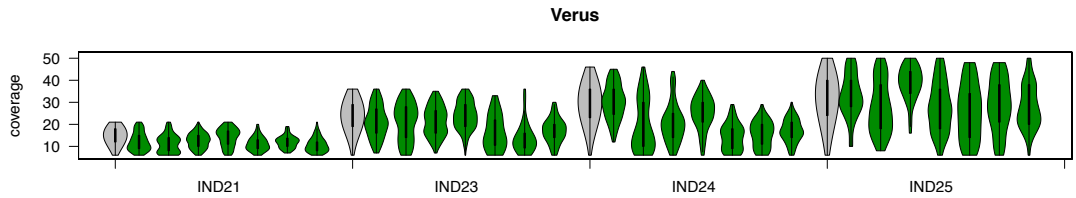


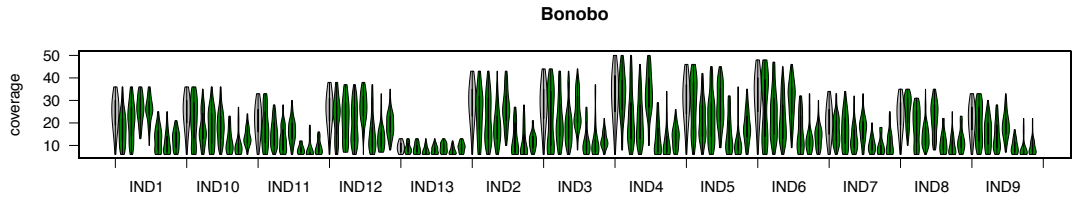
Figure S2: Enrichment of shared windows across NCD2 quantile bins of 0.05 between gorillas and orangutans. Observations based on empirical data are shown in blue, whereas results obtained in neutral simulations are shown in red.

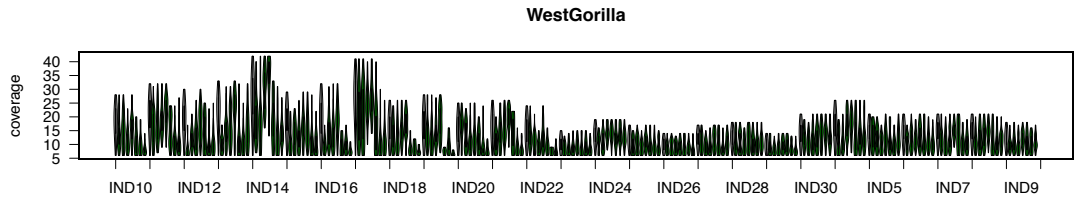

Ellioti
coverage
30
25
20
15
10
5
IND1
IND10
IND2
IND3
IND4
IND5
IND6
IND7
IND8
IND9


Schweinfurthii
coverage
50
40
30
20
10
IND11
IND12
IND13
IND14
IND15
IND16


Troglodytes
coverage
40
35
30
25
20
15
10
5
IND17
IND18
IND19
IND20


Verus
coverage
50
40
30
20
10
IND21
IND23
IND24
IND25


Bonobo
coverage
50
40
30
20
10
IND1
IND10
IND11
IND12
IND13
IND2
IND3
IND4
IND5
IND6
IND7
IND8
IND9


WestGorilla
coverage
40
35
30
25
20
15
10
5
IND10
IND12
IND14
IND16
IND18
IND20
IND22
IND24
IND26
IND28
IND30
IND5
IND7
IND9

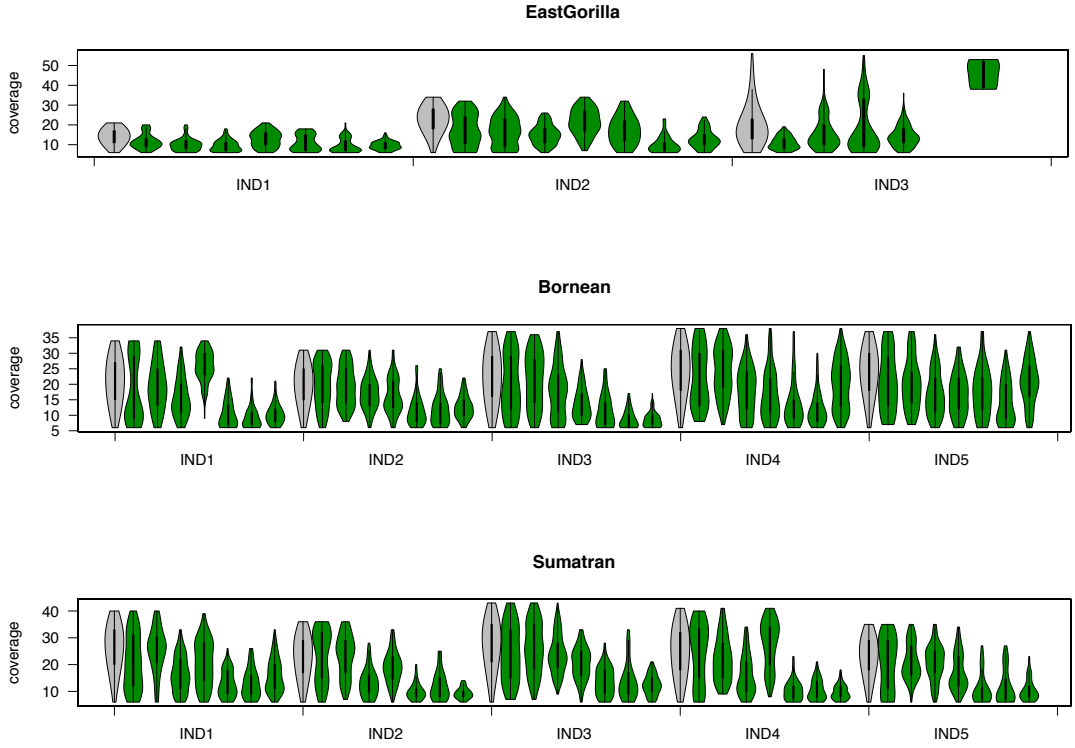


Figure S3: Coverage distribution (per individual) in seven candidate windows showing signatures of balancing selection in all great apes (shown in green). The genome-wide coverage is shown in grey and includes all positions in chromosomes 6, 12 and 17 (the three chromosomes on which these candidate windows were found). Coverage is computed across individuals in each species.

species	number of IS per quantile						% of windows kept after filtering					
	5%*	10%	15%	20%	25%	30%	5%	10%	15%	20%	25%	30%
Nigeria-Cameroon Chimpanzee	7	9	11	12	14	15	0.95	0.89	0.83	0.80	0.73	0.69
Central Chimpanzee	7	9	10	12	13	15	0.94	0.88	0.84	0.77	0.74	0.67
Eastern Chimpanzee	8	9	11	12	14	15	0.92	0.89	0.83	0.80	0.73	0.69
Western Chimpanzee	7	8	10	11	12	13	0.93	0.89	0.81	0.77	0.73	0.68
Bonobo	7	9	11	12	13	15	0.95	0.89	0.82	0.78	0.75	0.67
Western Lowlan Gorilla	10	13	15	17	19	21	0.94	0.88	0.83	0.78	0.73	0.67
Eastern Lowlan Gorilla	9	11	13	14	16	17	0.93	0.88	0.82	0.79	0.73	0.69
Bornean Orangutan	15	18	21	24	27	29	0.94	0.89	0.85	0.79	0.74	0.70
Sumatran Orangutan	15	18	22	25	28	31	0.94	0.90	0.84	0.79	0.74	0.68

Table S1: Number of *IS* per quantile in different species (left) and number of windows retained after filtering windows with low number of *IS*. *windows not considered

species	NCD variance after excluding windows (per quantile)						
	no-filter	5%	10%	15%	20%	25%	30%
Nigeria-Cameroon Chimpanzee	0.00077	0.00074	0.00071	0.00069	0.00069	0.00068	0.00067
Central Chimpanzee	0.00087	0.00083	0.00080	0.00078	0.00076	0.00076	0.00074
Eastern Chimpanzee	0.00092	0.00086	0.00085	0.00083	0.00082	0.00080	0.00080
Western Chimpanzee	0.00061	0.00058	0.00057	0.00056	0.00055	0.00055	0.00054
Bonobo	0.00043	0.00041	0.00040	0.00039	0.00038	0.00038	0.00037
Western Lowlan Gorilla	0.00063	0.00060	0.00058	0.00057	0.00056	0.00056	0.00055
Eastern Lowlan Gorilla	0.00053	0.00052	0.00051	0.00051	0.00051	0.00051	0.00051
Bornean Orangutan	0.00030	0.00028	0.00028	0.00027	0.00027	0.00027	0.00026
Sumatran Orangutan	0.00030	0.00028	0.00028	0.00027	0.00026	0.00026	0.00025

Table S2: NCD2 variance before and after filtering windows based on different quantiles of IS per species.

species	Number of windows analyzed
Nigeria-Cameroon Chimpanzee	1,354,649
Central Chimpanzee	1,330,282
Eastern Chimpanzee	1,361,391
Western Chimpanzee	1,313,518
Bonobo	1,346,134
Western Lowlan Gorilla	1,347,664
Eastern Lowlan Gorilla	1,348,318
Bornean Orangutan	1,356,400
Sumatran Orangutan	1,364,667

Table S3: Number of NCD windows considered in each species after filtering on the number of IS.

species pair	*% sharing*	*quantile*	*binomial P*
bonobo_abelli	0.0623	1	2,20E-21
bonobo_ellioti	0.1010	1	0
bonobo_gorilla	0.0702	1	9,65E-61
bonobo_graueri	0.0663	1	2,41E-78
bonobo_pygmaeus	0.0637	1	4,95E-31
bonobo_schweinfurthii	0.1047	1	0
bonobo_troglodytes	0.1059	1	0
bonobo_verus	0.0933	1	6,60E-273
ellioti_abelli	0.0702	1	3,15E-61
ellioti_gorilla	0.0763	1	5,28E-93
ellioti_gorilla	0.0752	20	1,58E-82
ellioti_graueri	0.0700	1	6,75E-79
ellioti_pygmaeus	0.0685	1	2,63E-53
ellioti_schweinfurthii	0.3855	1	0
ellioti_troglodytes	0.3860	1	0
ellioti_verus	0.2515	1	0
gorilla_abelli	0.0877	20	1,22E-195
gorilla_pygmaeus	0.0834	20	1,92E-145
graueri_abelli	0.0727	1	1,58E-85
graueri_abelli	0.0712	20	2,37E-60
graueri_gorilla	0.2283	1	0
graueri_pygmaeus	0.0914	20	0
pygmaeus_abelli	0.2458	1	0
schweinfurthii_abelli	0.0697	1	8,31E-56
schweinfurthii_gorilla	0.0755	1	2,70E-83
schweinfurthii_gorilla	0.0743	20	2,44E-71
schweinfurthii_graueri	0.0713	1	1,25E-93
schweinfurthii_pygmaeus	0.0685	1	2,64E-55
troglodytes_abelli	0.0681	1	2,47E-47
troglodytes_gorilla	0.0779	1	3,06E-105
troglodytes_graueri	0.0698	1	6,92E-78
troglodytes_pygmaeus	0.0652	1	3,34E-38
troglodytes_schweinfurthii	0.4776	1	0
verus_abelli	0.0627	1	5,66E-18
verus_gorilla	0.0719	1	8,24E-74
verus_graueri	0.0667	1	1,32E-75
verus_graueri	0.0667	20	1,43E-62
verus_pygmaeus	0.0647	20	5,58E-21
verus_schweinfurthii	0.2224	1	0
verus_troglodytes	0.2225	1	0

Table S4: Proportion of shared targets between pairs of species in significant quantiles (0.05) as defined by the bootstrapping analysis. The significant quantile and the p-value for the binomial (Bonferroni corrected) are indicated.

species	Ztf p-value	FDR	Category description	Genes
Nigeria-Cameroon Chimpanzee	0.005	0.00387	oxidoreductase activity	34
		0.00387	integral to membrane	309
		0.00387	antigen processing and presentation	ap3d1,hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,hla-f,hla-dqa1
		0.00387	MHC class II protein complex	hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,hla-dqa1
		0.00387	integral to lumenal side of endoplasmic reticulum membrane	hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,hla-f,hla-dqa1
		0.01012	clathrin coated endocytic vesicle membrane	hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,ldlr,hla-dqa1
		0.02393	MHC class II receptor activity	hla-dpa1,hla-drb1,hla-dqa1
Central Chimpanzee	0.005	0.00654	antigen processing and presentation	ap3d1,hla-dpb1,ap3b1,hla-drb5,hla-dra,hla-dqb1,hla-dqa1
		0.00654	MHC class II protein complex	hla-dpb1,hla-drb5,hla-dra,hla-dqb1,hla-dqa1
		0.00654	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	hla-dpb1,hla-dra,hla-dqb1,hla-dqa1
		0.02549	integral to membrane	303
Eastern Chimpanzee	0.005	0.01871	integral to membrane	307
		0.03105	extracellular region	98
Western Chimpanzee	0.005	0.00319	hemoglobin complex	hba1,hbd,hbe1,hbg2,hbq1
		0.00319	clathrin coated endocytic vesicle membrane	sh3gl2,hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,hla-dqa2,hla-dqa1
		0.00319	transport vesicle membrane	hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,ica1,ptprn2,hla-dqa2,hla-dqa1,slc30a8,scgn
		0.00319	MHC class II protein complex	hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,hla-dqb1,hla-dqa2,hla-dqa1
		0.00319	MHC class II receptor activity	hla-dpa1,hla-drb1,hla-dqa2,hla-dqa1
		0.00319	integral to lumenal side of endoplasmic reticulum membrane	hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,hla-dqa2,hla-dqa1
		0.00438	integral to membrane	282
		0.00438	antigen processing and presentation	hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,hla-dqb1,hla-dqa2,hla-dqa1
		0.00438	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	hla-dpb1,hla-dpa1,hla-dqb1,hla-dqa1
		0.03383	oxygen transporter activity	hba1,hbd,hbe1,hbg2,hbq1
		0.03383	collagen fibril organization	loxl2,ddr2,col11a1,adamts2,col1a1,col5a1,col5a3,col5a2,mkx,acan
		0.03490	ER to Golgi transport vesicle membrane	hla-dpb1,hla-dpa1,hla-drb1,hla-drb5,hla-dqa2,hla-dqa1
Bonobo	0.005	0.04096	MHC class II protein complex	hla-dpb1,hla-drb1,hla-dra,hla-dqa1
		0.04096	MHC class II receptor activity	hla-drb1,hla-dra,hla-dqa1
		0.04146	antigen processing and presentation	rab4a,hla-dpb1,hla-drb1,hla-f,hla-dra,hla-dqa1
		0.04240	integral to lumenal side of endoplasmic reticulum membrane	hla-dpb1,hla-drb1,hla-f,hla-dra,hla-dqa1
Western Lowland Gorilla	0.005	0.01933	integral to membrane	316
		0.04064	MHC class II protein complex	hla-dpa1,hla-drb1,hla-dqb1,hla-dqa1
Eastern Lowland Gorilla	0.005	0.00388	antigen processing and presentation	rab4a,hla-drb1,hla-dmb,hla-dra,hla-dqb1,hla-dqa2,hla-dqa1,ifng
		0.00388	transport vesicle membrane	cuzd1,hla-drb1,ica1,ptprn2,tmed10,hla-dra,hla-dqa2,hla-dqa1,slc30a8
		0.00388	MHC class II protein complex	hla-drb1,hla-dmb,hla-dra,hla-dqb1,hla-dqa2,hla-dqa1
		0.00388	MHC class II receptor activity	hla-drb1,hla-dra,hla-dqa2,hla-dqa1
		0.00388	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	hla-dmb,hla-dra,hla-dqb1,hla-dqa1
		0.04274	peroxidase activity	pxdn,hba1,ipcef1,mgst3,pxdnl,park7
Sumatran Orangutan	0.005	0.00636	integral to membrane	353
		0.00636	extracellular region	120
		0.00636	MHC class II protein complex	hla-dpb1,hla-dpa1,hla-drb1,hla-dqb1,hla-dqa1
		0.01501	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	hla-dpb1,hla-dpa1,hla-dqb1,hla-dqa1
		0.02121	clathrin coated endocytic vesicle membrane	sh3gl2,hla-dpb1,hla-dpa1,hla-drb1,hla-dqa1,tyrp1
		0.02121	MHC class II receptor activity	hla-dpa1,hla-drb1,hla-dqa1
Bornean Orangutan	0.005	0.01885	integral to membrane	342
		0.01915	pathogen associated molecular pattern dependent induction by symbiont of host innate immune response	tlr6,tlr2,tlr3
		0.02500	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	hla-dpb1,hla-dpa1,hla-dqb1,hla-dqa1
		0.02500	microglial cell activation involved in immune response	tlr6,tlr2,tlr3
		0.02500	phototransduction visible light	16
		0.02845	MHC class II protein complex	hla-dpb1,hla-dpa1,hla-dqb1,hla-dqa1
		0.04087	integral to lumenal side of endoplasmic reticulum membrane	hla-dpb1,hla-dpa1,sppl3,hla-dqa1,pkd2
		0.04698	detection of diacyl bacterial lipopeptide	tlr1,tlr6,tlr2

Table S5: Gene Ontology (GO) enriched categories for candidate targets of balancing selection in great apes.

6. Discussion

Balancing selection maintains advantageous genetic diversity through a variety of mechanisms (Charlesworth 2006, Andrés 2011, Key et al. 2014). Previous studies demonstrated that balancing selection in humans is responsible for shaping the evolution of genes involved in a variety of functions, particularly immune response to pathogens (Allison 1956, Pasvol et al. 1978, Fumagalli et al. 2009, Andrés et al. 2009). Nevertheless, only a few studies addressed how balancing selection affects the evolution of genes in our closest living relatives, the great apes. The lack of knowledge on targets of balancing selection in great apes represents a significant caveat in our understanding of how selection has shaped similarly (or differently) the evolution of different genes in humans and apes. This is of particular relevance since balancing selection is known to be able to act and maintain advantageous polymorphism for millions of years. In fact, with very few exceptions (Ferguson et al. 2012), the vast majority of studies of balancing selection in non-human primates focused on candidate gene approaches aiming to uncover trSNPs between humans and chimpanzees (Asthana et al. 2005, Cagliani et al. 2010, Cagliani et al. 2012), most of which are located in the MHC cluster (Asthana et al. 2005). Recently, Leffler et al. (2013) implemented a genome-wide survey that revealed the existence of six trans-species haplotypes maintained by balancing selection in humans and chimpanzees. Interestingly, none of these haplotypes contained protein-coding trSNPs, which led the authors to suggest a putative regulatory role for targets of long-term balancing selection. While this manuscript represented a big step forward in the field, the strategy adopted required the presence of at least two trSNPs in complete linkage in the two species, whereby ignoring cases where selection has maintained

only one trSNP. It remained unclear to what extent additional, single trSNPs exist in these two species. Further, while trSNPs represent strong evidence for the action of long-term balancing selection, they also likely represent a few of its targets in the genome. This is because selection may be younger than the common ancestor of these two species and, even if selection is ancestral, the selected polymorphism may have been lost in or even both populations due to demographic events or temporal changes in selective pressures. Thus, additional strategies hold the potential to identify targets of balancing selection, shared among species or species-specific, in the great apes.

This thesis aimed at uncovering regions of the genome with shared signatures of balancing selection in great apes, and the results here presented arise from the implementation of two different yet complementary strategies: in the first part, I describe instances of trSNPs in humans, chimpanzees and bonobos, which represent cases where balancing selection is shared since the common ancestor of the three species to the present-day populations of the three species; in the second part, I present a scan for signatures of balancing selection in the major great ape clades, and provide evidence that balancing selection is (and therefore selective pressures are) *shared* across different species. In addition, this part of thesis provides valuable information on the targets of balancing selection in each great species (no such analysis has been published to date).

I start by focusing on the most extreme cases, where balancing selection is shared and has continuously acted since the common ancestor of humans, chimpanzees and bonobos. We uncovered a total of 8 protein-coding trSNPs that are segregating for approximately 14 million years of independent evolution in the genomes of the three species. All of these trSNPs show the classical signatures of long-term balancing selection, such as segregating at intermediate allele frequencies and, most

importantly, defining haplotypes that show clustering by allele-type rather than by species. This last aspect is particularly relevant as it allows to exclude recurrent mutations in the different species, whereby providing (together with the other signatures of linked variation) overwhelmingly support for the influence of balancing selection maintaining these trSNPs .

The majority of the trSNPs belong to HLA genes in the MHC region, a finding that is in agreement with previous studies by others (Asthana et al. 2005, Leffler et al. 2013) and strengthens the evidence that the MHC cluster harbors the most extreme examples of balancing selection in the genome. Furthermore, we found a non-synonymous trSNP segregating in the gene *LAD1*, an autoantigen involved in cell adhesion that is responsible for linear IgA disease, which causes blistering of the skin (Ishiko et al. 1996, Marinkovich et al. 1996, Motoki et al. 1997, McKee et al. 2005). Interestingly, genes involved in cell adhesion have been previously reported as targets of balancing selection possibly due to denfense against pathogenic infections (Andrés et al. 2009, Fumagalli et al. 2009, 2011). An enticing possibility is that balancing selection targets an autoimmune gene because immune response should be effective in the fight against pathogens but at the same time moderate as to prevent self-recognition by the immune system (Ferrer-Admetlla et al. 2008). Discerning the biological basis of advantageous genetic diversity maintained by balancing selection in *LAD1* is an interesting problem that can be addressed by future work. Finally, it should be noted that while we present strong evidence for the rarity of trSNPs in humans, it is possible that additional examples exist. In particular, our strategy included the analysis of coding trSNPs, which hampers the detection of intergenic trSNPs that might play a role, for example, in regulating gene expression. Requiring the SNP to be present in three species (humans, chimpanzees and bonobos) reduces

their likelihood to be expected under neutrality, but also restricts our results to cases where all three species share the selective pressure. Finally, the strict filtering strategy we employed might have resulted on the exclusion of true trSNPs from the data.

As stated above, trSNPs represent striking cases of balancing selection shared across species. Nevertheless, it is possible for selection to act independently on the same loci in different species due to convergent evolution. Alternatively, even if balancing selection is maintained since the common ancestor of different species, it is possible for trSNPs to be lost in one species. In the second part of the thesis, I present our strategy to uncover and characterize targets of balancing selection in the genomes of great apes, focusing the analysis on patterns of shared signatures between species. We take advantage of the recent availability of great ape population data (Prado-Martínez et al. 2013), the lack of which has undoubtedly hampered the possibility to investigate this question in the past. Hence, this data makes it now possible not only to understand how balancing selection shapes the genomes of our closest living relatives, but also to paint a more complete picture on how similar selective pressures affect the evolution of different primates. Here, we focus on the targets of balancing selection. We start by showing that NCD2 (Bitarello et al. in preparation) is an powerful statistic to for our purposes, as it allows for the detection of targets of long-balancing selection in the genomes of all great apes.

We thus used NCD2 to analyze the genome. While the regions that result from this analysis are prime candidates to be targets of strong balancing selection, we need to care for possible alternative explanations for their unusual patterns of polymorphism. First, technically, it is possible that duplications in the genomes of great apes or, inversely, deletions in the human genome cause mapping problems that result in the

presence of a high number of artifact polymorphisms in particular windows. This possibility seems unlikely given our stringent filtering criteria and the fact that our top candidate windows have coverage levels comparable to those of the remainder of the genome. Biologically, and since we explicitly focus on windows showing high levels of polymorphism that segregate at intermediate frequencies, our candidate windows could in principle be considered to be compatible with relaxed evolutionary constraint. We show instead that candidate windows (those showing extremely low $NCD2_{value}$) are overlapping regions with functional relevance in the genome. In particular, we see a significant overlap with genic regions. When we focus on the regions that show signatures of balancing selection in several species, we observe also an enrichment in sites that are predicted to be involved in the regulation of gene expression. These results therefore allow ruling out relaxation of selective constraint and instead indicate balancing selection as the likely cause for the observed signatures. Overall, our data strongly supports NCD2 outlier windows as enriched for targets of balancing selection. Interestingly, GO analysis revealed that most significant categories include HLA genes (although many candidate windows overlap non-HLA genes), which again supports the idea that the MHC region is a preferential target of balancing selection.

Finally, we provide evidence that the amount of shared targets of balancing selection in different species is significantly higher than expected given shared ancestry, particularly if we consider comparison between more distantly related species like chimpanzee and gorilla or orangutan. These results are remarkable in that they sufggest that certain environmental pressures affect similarly the same loci in different species, and demonstrate the strength of our approach to uncover shared targets of balancing selection. In fact, we were able to uncover, in all great ape

species, signatures of selection in the exact same regions of the genes *HLA-DRB1*, *HLA-DRB5*, *HLA-DQB1*, *ATN1* and *GARNL4*/*RAP1GAP2*. All of these regions include positions that have been shown to affect the regulation of gene expression and, in the case of the three HLA genes, signatures of balancing selection are also seen in human populations (Bitarello et al. in preparation). Future detailed analysis of these regions is fundamental to understand the shared signatures we identified.

In conclusion, this thesis provides evidence for the action of long-term balancing selection affecting the genomes of humans and other great apes. We implement two complementary strategies and identify shared targets of balancing selection between different species, many of which include examples where the onset of selection is millions of years old, and some where it might even predate the split between some great ape species.

7. References

Abecasis, GR, A Auton, LD Brooks, MA DePristo, RM Durbin, RE Handsaker, HM Kang, GT Marth, GA McVean. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65.

Aboobaker, J, FT Wojnarowska, B Bhogal, MM Black. 1991. Chronic bullous dermatosis of childhood--clinical and immunological features seen in African patients. Clin Exp Dermatol 16:160-164.

Alkan, C, JM Kidd, T Marques-Bonet, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41:1061-1067.

Alkhatib, C, C Combadiere, CC Broder, et al. 1996. CC CKR5: A RANTES, MIP-1a, MIP1-b receptor as a fusion cofactor for macrophage-trofic HIV-I. Science 272(5270):1955-58.

Allison, AC. 1956. The sickle-cell and haemoglobin C genes in some African populations. Ann Hum Genet 21:67-89.

Alonso S, S López, N Izagirre, C de la Rúa. 2008. Overdominance in the human genome and olfactory receptor activity. Mol Biol Evol 25:997-1001

Anders, S, W Huber. 2010. Differential expression analysis for sequence count data. Genome Biol 11:R106.

Andrés, AM. 2011. Balancing Selection in the Human Genome. Encyclopedia of Life Sciences (eLS). Chichester: John Wiley & Sons Ltd.

Andrés, AM, MY Dennis, WW Kretzschmar, 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, MJ Hubisz, A Indap, et al. 2009. Targets of balancing selection in the human genome. Mol Biol Evol 26:2755-2764.

Assinger, A. 2014. Platelets and infection: an emerging role of platelets in viral infection. Front Immunol. 5:649.

Asthana, S, S Schmidt, S Sunyaev. 2005. A limited role for balancing selection. Trends Genet 21:30-32.

Bamshad, MJ, S Mummidi, E Gonzalez, et al. 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci USA 99:10539-10544.

Barton, NH, AM Etheridge. 2004. The effect of selection on genealogies. Genetics 166: 1115–1131.

Bernhard, W, K Barreto, S Raithatha, I Sadowski. 2013. An upstream YY1 binding site on the HIV-1 LTR contributes to latent infection. PLos One http://dx.doi.org/10.1371/journal.pone.0077052.

Bird, AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499-1504.

Bitarello, B, C de Filippo, JC Teixeira, et al. The targets of long-term balancing selection in human populations. In Prep.

Boyle, AP, EL Hong, M Hariharan, et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. Gen Res 22(9):1790-1797.

Bubb, KL, D Bovee, D Buckley, et al. 2006. Scan of human genome reveals no new loci under ancient balancing selection. Genetics 173: 2165–2177.

Cagliani, R, M Fumagalli, M Biasin, L Piacentini, S Riva, U Pozzoli, MC Bonaglia, N Bresolin, M Clerici, M Sironi. 2010. Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene. Hum Genet 128:577- 588.

Cagliani, R, FR Guerini, M Fumagalli, et al. 2012. A trans-specific polymorphism in ZC3HAV1 is maintained by long-standing balancing selection and may confer susceptibility to multiple sclerosis. Mol Biol Evol 29:1599-1613.

Charlesworth, B, M Nordborg, D Charlesworth. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res 70:155-174.

Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64.

Cheng, Z, M Ventura, X She, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88-93.

Christensen, GL, IP Ivanov, SP Wooding, et al. 2006. Identification of polymorphisms and balancing selection in the male infertility candidate gene, ornithine decarboxylase antizyme 3. BMC Med Genet 7:27

Clark, AG. 1997. Neutral behavior of shared polymorphism. Proc Natl Acad Sci USA 94:7730-7734.

Coull, JJ, F Romerio, JM Sun, JL Volker, KM Galvin, et al. 2000. The human factors YY1 and LSF repress the human immunodeficiency virus type 1 long terminal repeat via recruitment of histone deacetylase 1. J Virol 74: 6790-6799.

Cutrera, AP, EA Lacey. 2007. Trans-species polymorphism and evidence of selection on class II MHC loci in tuco-tucos (Rodentia: Ctenomyidae). Immunogenetics 59:937-948.

Darwin, C, AR Wallace. 1858. Proceedings of Linnean Society of London 3, 45.

DeGiorgio, M, K Lohmueller, R Nielsen. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet 10(8):e1004561

Denguezli, M, B Ben Nejma, R Nouira, S Korbi, R Bardi, K Ayed, AS Essoussi, B Jomaa. 1994. [Iga linear bullous dermatosis in children. A series of 12 Tunisian patients]. Ann Dermatol Venereol 121:888-892.

Derrien, T, J Estelle, S Marco Sola, DG Knowles, E Raineri, R Guigo, P Ribeca. 2012. Fast computation and applications of genome mappability. PLoS One 7:e30377.

Drmanac, R, AB Sparks, MJ Callow, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327(5961)78-81.
Ewing, G, J Hermisson. 2010. MSMS: A Coalescent Simulation Program Including Recombination, Demographic Structure, and Selection at a Single Locus. Bioinformatics doi: 10.1093/bioinformatics/btq322

Fan, WM, M Kasahara, J Gutknecht, D Klein, WE Mayer, M Jonker, J Klein. 1989. Shared class II MHC polymorphisms between humans and chimpanzees. Hum Immunol 26:107-121.

Fasquelle, C, A Sartelet, W Li, et al. 2009. Balancing selection of a frame-shift mutation in the MRC2 gene accounts for the outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. PLoS Genet 5:e1000666.

Favorov, A, L Mularoni, L Cope, et al. 2012. Exploring massive, genome scale datasets with the GenometriCorr package. PLos Comp Bio http://dx.doi.org/10.1371/journal.pcbi.1002529

Ferguson, W, S Dvora, RW Fikes, AC Stone, S Boissinot. 2012. Long-term balancing selection at the antiviral gene OAS1 in Central African chimpanzees. Mol Biol Evol 29:1093-1103.

Ferrer-Admetlla, A, E Bosch, M Sikora, et al. 2008. Balancing selection is the main force shaping the evolution of innate immunity genes. J Immunol 181:1315-1322.

Fischer, A, K Prüfer, JM Good, M Halbwax, V Wiebe, C Andre, R Atencia, L Mugisha, SE Ptak, S Paabo. 2011. Bonobos fall within the genomic variation of chimpanzees. PLoS One 6:e21605.

Fumagalli, M, R Cagliani, U Pozzoli, S Riva, GP Comi, G Menozzi, N Bresolin, M Sironi. 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res 19:199-212.

Fumagalli, M, U Pozzoli, R Cagliani, et al. 2010. Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach. PLoS Genet 6:e1000849

Fumagalli, M, M Sironi, U Pozzoli, A Ferrer-Admetlla, L Pattini, R Nielsen. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet 7:e1002355.

Gendzekhadze, K, PJ Norman, L Abi-Rached, T Graef, AK Moesta, Z Layrisse, P Parham. 2009. Co-evolution of KIR2DL3 with HLA-C in a human population retaining minimal essential diversity of KIR and HLA class I ligands. Proc Natl Acad Sci USA 106:18692-18697.

Germain, RN. 1994 MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. Cell 76:287-299

Gigord, LD, MR Macnair, A Smithson. 2001. Negative frequency-dependent selection maintains a dramatic flower color polymorphism in the rewardless orchid Dactylorhiza sambucina (L.) Soo. Proc Natl Acad Sci USA 98:6253-6255.

Gillespie, JH. 1978. A general model to account for enzyme variation in natural populations. V. The SAS--CFF model. Theor Popul Biol 14:1-45.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Graser, R, C O'HUigin, V Vincek, A Meyer, J Klein. 1996. Trans-species polymorphism of class II Mhc loci in danio fishes. Immunogenetics 44:36-48.

Griffiths, RC, S Tavare. 1994. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci 344:403-410.

Hamm, D, BS Mautz, MF Wolfner, CF Aquadro, WJ Swanson. 2007. Evidence of amino acid diversity-enhancing selection within humans and among primates at the candidate sperm-receptor gene PKDREJ. Am J Hum Genet 81:44-52.

Harding, CV, HJ Geuze. 1993. Antigen processing and intracellular traffic of antigens and MHC molecules. Curr Opin Cell Biol 5:596-605

Hartl, D, AG Clark. 1989. Principles of population genetics. Sunderland, MA: 2nd ed.

Hintze, JL, RD Nelson. 1998. Violin plots: a box plot-density trace synergism. The American Statistician 52(2):181-184.

Hodgkinson, A, A Eyre-Walker. 2010. The genomic distribution and local context of coincident SNPs in human and chimpanzee. Genome Biol Evol 2:547-557.

Hodgkinson, A, A Eyre-Walker. 2011. Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12:756-766.

Hodgkinson, A, E Ladoukakis, A Eyre-Walker. 2009. Cryptic variation in the human mutation rate. PLoS Biol 7:e1000027.

Hudson, RR, M Kreitman, M Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.

Hudson, RR, NL Kaplan. 1988. The coalescent process in models with selection and recombination. Genetics 120:831-840.

Hudson, RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Hughes, AL, M Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170.

Hughes, AL, M Nei. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci USA 86:958-962.

Hughes, AL, M Yeager. 1998. Natural selection at major histocompatibility complex loci in vertebrates. Annu Rev Genet 32:415-35

Hughes KA, AE Houde, AC Price, FH Rodd. 2013. Mating advantage for rare males in wild guppy populations. Nature 503:108-110.

Ishiko, A, H Shimizu, T Masunaga, T Hashimoto, M Dmochowski, F Wojnarowska, BS Bhogal, MM Black, T Nishikawa. 1996. 97-kDa linear IgA bullous dermatosis (LAD) antigen localizes to the lamina lucida of the epidermal basement membrane. J Invest Dermatol 106:739-743.

Johnson, PL, I Hellmann. 2011. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. Genome Biol Evol 3:842-850.

Johnston, SE, J Gratten, C Berenos, et al. 2013. Life history trade-offs at a single locus maintain sexually selected genetic variation. Nature 502:93-95.

Kent, WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12:656-664.

Key, FM, JC Teixeira, C de Filippo, AM Andrés. 2014. Advantageous diversity maintained by balancing selection in humans. Curr Opin Genet Dev 29C:45-51.

Kikkawa, EF, TT Tsuda, D Sumiyama, et al. 2009. Trans-species polymorphism of the Mhc class II DRB-like gene in banded penguins (genus Spheniscus). Immunogenetics 61:341-352.

Kim, MS, SM Pinto, D Getnet, et al. 2014. A draft map of the human proteome. Nature 509:575-581.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kircher, M, S Sawyer, M Meyer. 2012. Double indexing overcomes inaccuracies in

multiplex sequencing on the Illumina platform. Nucleic Acids Res 40:e3.

Kircher, M, U Stenzel, J Kelso. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol 10:R83.

Klein, J, Y Satta, C O'HUigin, N Takahata. 1993. The molecular descent of the major histocompatibility complex. Annu Rev Immunol 11:269-295.

Kofler, R, C Schlöterer. 2012. Gowinda: an unbiased analysis of gene set enrichment for genome-wide association studies. Bioinformatics. doi: 10.1093/bioinformatics/bts315

Koide R, T Ikeuchi, O Onodera, et al. 1994. Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nature Genetics 6(1):9–13.

Kroyman, J, T Mitchell-Olds. 2005. Epistasis and balanced polymorphism influencing complex trait variation. Nature 435:95-98.

Lappalainen, T, M Sammeth, MR Friedlander, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501:506-511.

Lawlor, DA, FE Ward, PD Ennis, AP Jackson, P Parham. 1988. HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. Nature 335:268-271.

Leffler, EM, Z Gao, S Pfeifer, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339:1578-1582.

Li, H, R Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754-60. doi: 10.1093/bioinformatics/btp324.

Loisel, DA, MV Rockman, GA Wray, J Altmann, SC Alberts. 2006. Ancient polymorphism and functional variation in the primate MHC-DQA1 5' cis-regulatory region. Proc Natl Acad Sci USA 103:16331-16336.

Lynch, M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci USA 104 Suppl 1:8597-604.

Margolis, DM, M Somasundaran, MR Green. 1994. Human transcription factor YY1 Represses human immunodeficiency virus type 1 transcription and virion production. J Virol 68(2):905-910.

Marinkovich, MP, TB Taylor, DR Keene, RE Burgeson, JJ Zone. 1996. LAD-1, the linear IgA bullous dermatosis autoantigen, is a novel 120-kDa anchoring filament protein synthesized by epidermal cells. J Invest Dermatol 106:734-738.

Mayer, WE, M Jonker, D Klein, P Ivanyi, G van Seventer, J Klein. 1988. Nucleotide sequences of chimpanzee MHC class I alleles: evidence for trans-species mode of evolution. EMBO J 7:2765-2774.

McKee, PH, E Calonje, SR Granter. 2005. Pathology of the skin : with clinical correlations / [edited by] Phillip H. McKee, Eduardo Calonje, Scott R. Granter. Edinburgh: Philadelphia Elsevier Mosby.

McKenna, A, M Hanna, E Banks, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303.

Meyer, M, M Kircher. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc 2010:pdb prot5448.

Mendez, FL, JC Watkins, MF Hammer. 2013. Neandertal origin of genetic variation at the cluster of OAS immunity genes. Mol Biol Evol 30(4):798-801

Monia, K, K Aida, K Amel, Z Ines, F Becima, KM Ridha. 2011. Linear IgA bullous dermatosis in tunisian children: 31 cases. Indian J Dermatol 56:153-159.

Mosher, DS, P Quignon, CD Bustamante, et al. 2007. A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs. PLoS Genet 3:e79.

Motoki, K, M Megahed, S LaForgia, J Uitto. 1997. Cloning and chromosomal mapping of mouse ladinin, a novel basement membrane zone component. Genomics 39:323-330.

Muehlenbachs, A, M Fried, J Lachowitzer, TK Mutabingwa, PE Duffy. 2008. Natural selection of FLT1 alleles and their association with malaria resistance in utero. Proc Natl Acad Sci USA 105:14488-14491.

Muirhead, CA, NL Glass, M Slatkin. 2002. Multilocus self-recognition systems in fungi as a cause of trans-species polymorphism. Genetics 161:633-641.

Nagafuchi S, H Yanagisawa, K Sato, et. al 1994. Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. Nature Genetics 6 (1):14–8.

Nei, M. 1975. Molecular population genetics and evolution. Front Biol 40:I-288

Nordborg, M. 1997. Structured coalescent processes on different time scales. Genetics 146: 1501–1514.

Olendorf, R, FH Rodd, D Punzalan, et al. 2006. Frequency-dependent survival in natural guppy populations. Nature. 441:633-636.

Pasvol, G, DJ Weatherall, RJ Wilson. 1978. Cellular mechanism for the protective effect of haemoglobin S against P. falciparum malaria. Nature 274:701-703.

Pereira, LA, K Bentley, A Peeters, et al. 2000. A compilation of cellular transcription

factor interactions with the HIV-1 LTR promoter. Nucleic Acids Res 28:663-668.

Prado-Martinez, J, PH Sudmant, JM Kidd, et al. 2013. Great ape genetic diversity and population history. Nature 499:471-475.

Prüfer, K, K Munch, I Hellmann, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486:527-531.

Prugnolle, F, A Manica, M Charpentier, JF Guégan, V Guernier, F Balloux. 2005. Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15(11):1022-7.

Ranasinghe, S, S Cutler, I Davis, et al. 2013. Association of HLA-DRB1-restricted $CD4^+$ T cell responses with HIV immune control. Nature Medicine 19:930-933.

Rasmussen MD, MJ Hubisz MJ, I Gronau, A Siepel. 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genet 10:e1004342

Roux, C, M Pauwels, MV Ruggiero, D Charlesworth, V Castric, X Vekemans. 2013. Recent and ancient signature of balancing selection around the S-locus in Arabidopsis halleri and A. lyrata. Mol Biol Evol 30:435-447

Rozen, S, H Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365-386.

Sabeti, PC, DE Reich, JM Higgins, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature. 419(6909):832-7.

Sabeti, PC, SF Schaffner, B Fry, et al. 2006. Positive natural selection in the human lineage. Science 312(5780):1614-20

Scheet, P, M Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629-644.

Schierup, MH, AM Mikkelsen, J Hein. 2001. Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. Genetics 159:1833-1844.

Ségurel, L, EE Thompson, T Flutre, et al. 2012. The ABO blood group is a transspecies polymorphism in primates. Proc Natl Acad Sci U S A 109:18493-18498.

Semple, JW, JE Italiano Jr, J Freedman. 2011. Platelets and the immune condition. Nat Rev Immunol. 11:264-274.

Sudmant, PH, J Huddleston, CR Catacchio, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. Gen Res 23(9):1373-82.

Sutton, JT, BC Robertson, CE Grueber, JA Stanton, IG Jamieson. 2013. Characterization of MHC class II B polymorphism in bottlenecked New Zealand saddlebacks reveals low levels of genetic diversity. Immunogenetics 65:619-633.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Takahata, N, M Nei. 1990. Allelic genealogy under overdominant and frequencydependent selection and polymorphism of major histocompatibility complex
loci. Genetics 124:967-978.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor Popul Biol 26:119-164.

Teixeira, JC, C de Filippo, A Weihmann, et al. 2015. Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees and bonobos. Mol Biol Evol 32(5):1186-96.

von Hundelshausen, P, C Weber. 2007. Platelets as immune cells: bridging inflammation and cardiovascular disease. Circ Res 100(1):27-40.

Ward, LD, M Kellis. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337(6102):1675-8

Weir, BS, CC Cockerham. 1984. Estimating F-Statistics for the Analysis of Population-Structure. Evolution 38:1358-1370.

Williamson, S, A Fledel-Alon, CD Bustamante. 2004. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. Genetics 168: 463–475.

Wiuf, C, K Zhao, H Innan, M Nordborg. 2004. The probability and chromosomal extent of trans-specific polymorphism. Genetics 168:2363-2372.

Wooding, S, UK Kim, MJ Bamshad, J Larsen, LB Jorde, D Drayna. 2004. Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. Am J Hum Genet 74:637-646.

Wooding, S, AC Stone, DM Dunn, S Mummidi, LB Jorde, RK Weiss, S Ahuja, MJ Bamshad. 2005. Contrasting effects of natural selection on human and chimpanzee CC chemokine receptor 5. Am J Hum Genet 76:291-301.

Wright, S. 1939. The Distribution of Self-Sterility Alleles in Populations. Genetics 24:538-552.

Wu, J, SJ Saupe, NL Glass. 1998. Evidence for balancing selection operating at the het-c heterokaryon incompatibility locus in a group of filamentous fungi. Proc Natl Acad Sci 95:12398-12403.

Yeaman, MR. 2014. Platelets: at the nexus of antimicrobial defence. Nat Rev Microbiol. 12:426-437.

Acknowledgments

First and foremost, I am thankful to my supervisor, Aida Andrés for her guidance, support and trust. Like most people when going through their PhD studies there were quite stressful times and Aida was always friendly and willing to help. Even when that meant that I would stop by her office every 20 minutes to share a new result (or, more often, a new problem in the analysis). I learned a lot studying with you and it makes me truly sad that we will not be able to work together daily from now on.

I thank my entire group, the Genetic Diversity and Selection! You guys are amazing and without you this would not be possible to accomplish. Mostly, I want to thank Cesare de Filippo and Joshua Schmidt, whom apart from being great friends, worked closely with me in the two projects presented in this work. Your contribution is only diminished by your friendship. I will miss you both dearly. Felix M. Key for being my "older brother" in the group. We made a friendship that will last for life and I'll miss deeply our (sometimes heated) debates.

I am also thankful to Svante Pääbo for leading quite exemplarily the department of evolutionary genetics. Success is much more tangible because the working environment is absolutely great.

Thank you to all the people in the department but particularly to Michael Dannemann and Leonardo Arias for your friendship and great humor. I will miss you both a lot!

Finally, I am thankful for all the people I have in my life outside work. Mostly my wonderful parents who are always there for me no matter the circumstances. Knowing this makes you happy allowed me to get going when things were not fun. I love you both.

My family and my friends.

Finally, thank you Sara. I cannot express in words what your love and support means to me. You always stood by my side and the geographical distance between us provides no barrier to our love.