

# Long-Term Balancing Selection in *LAD1* Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and Bonobos

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

<sup>1</sup>Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

<sup>2</sup>Department of Integrative Biology, University of California, Berkeley

<sup>3</sup>International Center for Insect Physiology and Ecology, Nairobi, Kenya

<sup>4</sup>Clinique vétérinaire du Dr. Jacquemin, Maisons-Alfort, France

<sup>5</sup>Lola Ya Bonobo sanctuary, Kinshasa, Democratic Republic Congo

<sup>6</sup>Réserve Naturelle Sanctuaire à Chimpanzés de Tchimpounga, Jane Goodall Institute, Pointe-Noire, Republic of Congo

<sup>†</sup>These authors contributed equally to this work.

**\*Corresponding author:** E-mail: aida\_andres@eva.mpg.de.

**Associate editor:** John Novembre

**Data deposition:** The exome sequence capture data reported in this paper have been deposited in the European Nucleotide Archive (accession no. PRJEB8350). The variant calls can be accessed at <http://cdna.eva.mpg.de/primate/exomes/VCFs>.

## 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 My of independent evolution. Although the majority of these trSNPs were found in three genes of the major histocompatibility locus 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*.

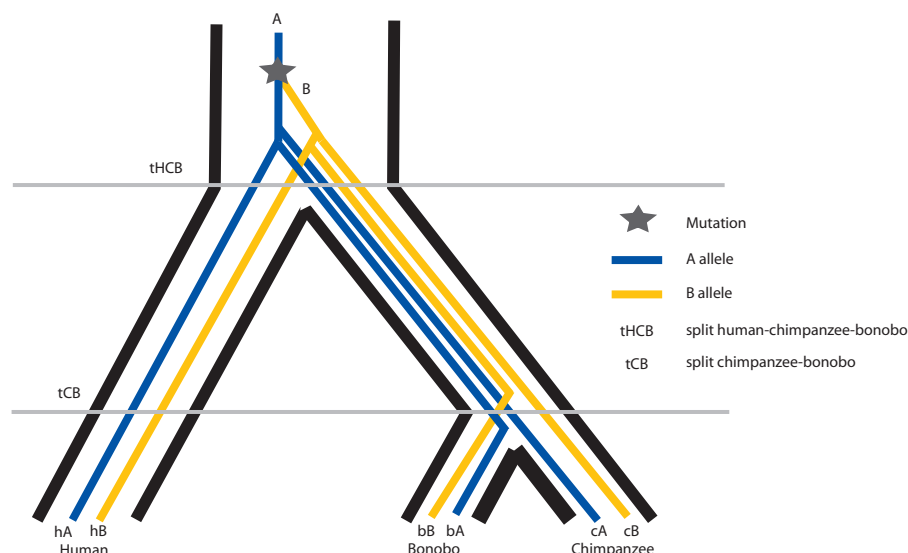
**Key words:** balancing selection, trans-species polymorphism, *LAD1*, great apes.

## 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; Gignord 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; fig. 1). In species with old enough divergence time trSNPs are rare under neutrality and are hallmarks of balancing selection (Charlesworth et al. 1997; Clark 1997; Wiuf et al. 2004).



**FIG. 1.** Schematic representation of a possible genealogy leading to a trSNP in human, chimpanzee, and bonobo.

The assumption that trSNPs are very rare in humans combined with the absence of unbiased genome-wide polymorphism data sets in other great ape species resulted in few trSNPs 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 trSNPs 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 trSNPs 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 trSNP (rs12088790) maintained by long-term balancing selection in the gene *LAD1* (*ladinin-1*).

## Results

### 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 SI, Supplementary Material online), 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 Ma (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 SI, Supplementary Material online). Conversely, a neutral trSNP 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  $P_{HC} = 1.6 \times 10^{-8}$  of being polymorphic in chimpanzees too (see also supplementary information SI, Supplementary Material online). 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  $4.0 \times 10^{-10}$ . 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  $5.0 \times 10^{-5}$  neutral trSNPs in the three species. We note that these are actually overestimates, because 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 SI, Supplementary Material online. 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.

### Identification of trSNPs

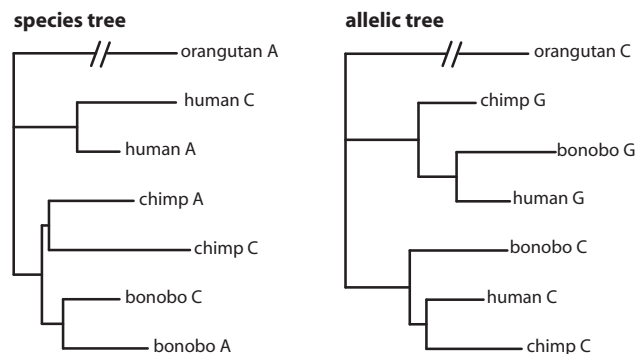
We sequenced the exomes of 20 Yoruba humans, 20 central chimpanzees (*Pan troglodytes troglodytes*), and 20 bonobos

(*Pan paniscus*) to an average coverage of approximately  $18\times$  in each individual (data are 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 approximately 33.54% in bonobo, approximately 49.29% in chimpanzee, and approximately 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 when using the 24-mer 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 SII, Supplementary Material](#) online). 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 2010; 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 (~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 ([fig. 2](#)) because the most recent common ancestor of the genomic segment containing a human SNP falls in the human branch, predating (backward in time) the coalescence of lineages from the different species (see previous section). On the contrary, trSNPs create local genealogies that cluster by allele ([fig. 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.

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 ( $P_{\text{allelic}} > 0.90$ ) ([supplementary table S4, Supplementary Material](#) online); these shSNPs, all of which are present in dbSNP build 138, were



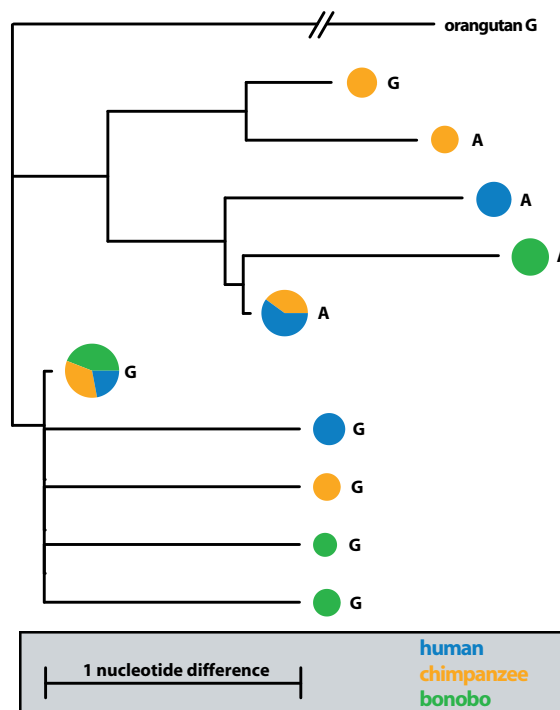
**Fig. 2.** Examples of a species tree and an allelic tree using six haplotypes, one per species and allele. Each Neighbor-joining tree is computed on a 500 bp region around a shSNP in our data set for the genes *TXNDC2* (species trees) and *HLA-DQA1* (allelic tree).

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 ([supplementary table S4, Supplementary Material](#) online). 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.

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 (i.e., 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 SII and table S4, Supplementary Material](#) online). 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.

### 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



**Fig. 3.** Neighbor-joining tree of *LAD1*. The tree was constructed using a 350 bp region as described in Materials and 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).

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 pseudorecurrent mutations where to investigate the neutral probability of an allelic tree (see Materials and Methods, [supplementary information SIII](#) and [table S1](#), [Supplementary Material](#) online, for details). We found that, as expected, the FDR is inversely proportional to the length of an allelic tree ([supplementary information SIII](#), [Supplementary Material](#) online); 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 ([supplementary table S1](#), [Supplementary Material](#) online).

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 100 bp and four 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](#)).

### 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 SIV](#), [Supplementary Material](#) online) in the genes containing our four non-HLA candidate trSNPs (*LY9*, *LAD1*, *SLCO1A2*, and *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: 1) “ALL”—the entire genic region; 2) “coding”—only their coding exonic sequence; 3) “500 bp”—the 500 bp surrounding the trSNP; and 4) the “length of allelic tree” ([supplementary table S4](#), [Supplementary Material](#) online). 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 ([supplementary table S4](#), [Supplementary Material](#) online). 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 nonsignificant in chimpanzee (see [tables 1](#) and [2](#); [supplementary table S4](#) and [information SIV](#), [Supplementary Material](#) online). The weaker signal in chimpanzee is likely due to this species' larger effective population size (Prado-Martinez 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



**Table 1.** Comparison of Different Signatures in Candidate trSNPs and Genes.

chr:position	Gene	Tree bp (FDR%)	#SNPs (H;C;B)	#FDs (H;C;B)	PtoD (p)				MAF		
					H	C	B	3spp	H	C	B
1:160788067 <sup>a</sup>	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 <sup>a</sup>	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 <sup>a</sup>	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 <sup>a</sup>		250 (0.0)	(19;14;13)	(0;0;0)					0.500	0.350	0.050
6:32609271 <sup>a</sup>		750 (0.0)	(25;23;24)	(0;0;0)					0.475	0.400	0.050
6:33052736 <sup>a</sup>	HLA-DPB1	1,000 (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		1,000 (0.0)	(8;10;11)	(0;0;0)					0.300	0.450	0.350
6:33052768		1,000 (0.0)	(8;10;11)	(0;0;0)					0.300	0.475	0.350
12:21453466	SLC1A2	1,000 (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 <sup>a</sup>	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

NOTE.—H, human; C, chimpanzee; B, bonobo. The estimated length of the allelic trees (and respective FDR), the number of polymorphisms (#SNPs) and fixed differences (#FDs) in the allelic tree, the *PtoD* ratios for the whole gene and MAFs 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 italic.

<sup>a</sup>Nonsynonymous trSNP.

**Table 2.** *PtoD* Ratios Calculated in the Gene *LAD1* (with the corresponding percentile in the empirical distribution in parenthesis).

<i>PtoD</i>	Human	Chimpanzee	Bonobo
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)
500 bp	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)	

NOTE.—For human, we present the *PtoD* ratio obtained in the human–bonobo comparison, which is very similar to the human–chimpanzee comparison.

presence of a single additional SNP in humans (in such a small region). *OAS1* shows significant excess of polymorphism only in chimpanzee.

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 and 2 and supplementary table S4, Supplementary Material online). 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 182 bp 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

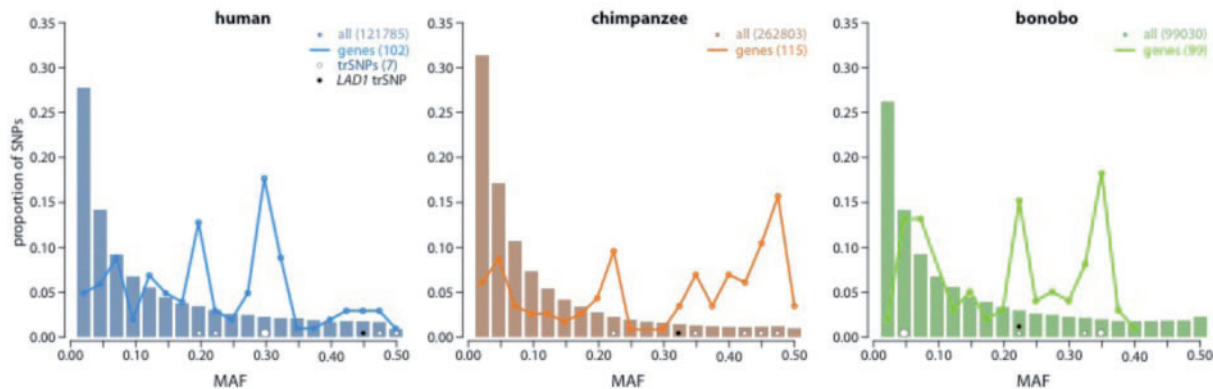
*SLC1A2*, *LY9* or *OAS1* have been under balancing selection, but conservatively we focus on *LAD1*, *HLA-C*, *HLA-DQA1*, and *HLA-DPB1* as our final set of trSNPs.

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 data set (supplementary tables S2 and S3 and figs. S6 and S7, Supplementary Material online). *LAD1* is the least polymorphic of the four genes, which is not surprising as the remaining trSNPs fall in HLA genes.

### 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 toward intermediate-frequency alleles, in all species (Mann–Whitney *U* test [MW-*U*]  $P < 4 \times 10^{-10}$ ; fig. 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 nonsignificant), 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 (MW-*U* test  $P \leq 0.046$ , table 3), including *LAD1*.



**Fig. 4.** Folded SFS of trSNPs and other SNPs in the genes. The x axis represents the 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.

**Table 3.** P-values (MW-U test) for Excess of Intermediate-Frequency Alleles Comparing the SFS of the Genes to the Genome-Wide SFS.

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

The trSNP in LAD1, which is a missense polymorphism, is at intermediate frequency in all three species (table 1 and fig. 4): minor allele frequency (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 nonsynonymous variants: In the upper 1.9% quantile for human, in the 8.6% for chimpanzee, and in the 23.8% in bonobo.

When we investigate the 1000 Genomes data set (Abecasis et al. 2012), which contains both coding and noncoding 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 (supplementary table S6, Supplementary Material online). The trSNP is itself present in all these populations throughout the world: At intermediate frequency in all African populations ( $31\% < \text{MAF} < 48\%$ ) and at lower frequency ( $\text{MAF} < 8\%$ ) 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 ( $30\% < \text{MAF} < 50\%$ ), the  $F_{ST}$  of rs12088790 is not in the upper tail ( $0.114 < P < 0.310$ —supplementary table S9, Supplementary Material online). The polymorphism is thus

shared across human populations, although its frequency shows certain differences among human groups.

### Balancing Selection in LAD1

LAD1 (*ladinin 1*) spans 18,704 bp and is composed of ten exons. We obtained a total of 1,213 bp 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  $\rightarrow$  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.

### 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, 2005; Muehlenbachs et al. 2008; Andrés et al. 2009, 2010), but here we focus on selection that is old, strong, constant, and shared across lineages, and that results in trSNPs. 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 world 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  $5.0 \times 10^{-5}$  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 eight trSNPs that pass our filtering criteria and that cluster by allelic tree, with an extremely low probability under recurrent mutation. The loci containing these eight SNPs show, in addition, the excess of polymorphism expected under long-term balancing selection. Seven trSNPs are located in *HLA* genes (*HLA-DQA1*, *HLA-C*, and *HLA-DPB1*) and one is a nonsynonymous 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 trSNP 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 nonsignificant 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 C, Key FM, Ghirotto S, Benazzo A, Meneu JR, Weihmann A, NISC Comparative Sequence Program, Parra G, Green ED, and Andrés AM et al., submitted). 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 nonsynonymous 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 (Ségurel 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-Martinez et al. 2013). Besides rs12088790 (which in that data set 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 kb downstream rs12088790 (see [supplementary information SVI, Supplementary Material](#) online). 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 SVI, Supplementary Material](#) online). 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 nonsynonymous 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, 2011; 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 auto-antibodies 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 autoimmunity 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.

## Materials and Methods

### DNA Samples and Sequencing

We performed whole-exome capture and high-coverage sequencing of 20 humans, 20 central chimpanzees (*P. troglodytes troglodytes*), and 20 bonobos (*P. 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 Puregene Tissue Kit (Qiagen), sheared to a size range of 200–300 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 50 Mb 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 76 bp.

### 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)  $\geq 25$  and mapping outside of known segmental duplications in the three species were considered for further analysis. Specifically, the average coverage for each individual is 18.9 $\times$  in human, 17.9 $\times$  in chimp, and 17.9 $\times$  in bonobo.

### 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 was less than 8 or greater than 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 data sets above [28–30];
- The quality score of the call was less than 50;
- There was evidence of strand bias (SB > 0);
- The genotype quality was less than 10 in all individuals carrying the alternative allele;
- The SNP was located within 3 bp of a homopolymer with a minimum length of 5 bp;
- The SNP was located within 5 bp up- and downstream of an insertion or deletion (indel) polymorphism or substitution with the human reference genome.



## Shared SNPs as trSNPs

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 (i.e., positions that have a 24-mer mappability score lower than 1, as defined by the CRG Alignability track in the UCSC browser [Derrien et al. 2012]) (see Results).

## 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 trSNP 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” (fig. 2). The trees were inferred in windows of different lengths (from 100 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 trSNPs only shSNPs that show an allelic tree with probability ( $P_{\text{allelic}}$ ) > 0.9 in a window of at least 100 nt.

We adopted a simple resampling approach to calculate  $P_{\text{allelic}}$  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).  $P_{\text{allelic}}$  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 (FDRs) under recurrent mutation and based on our empirical data set. For each observed allelic tree lengths (supplementary table S1, Supplementary Material online), 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  $P_{\text{allelic}}$  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  $P_{\text{allelic}} > 0.9$  (i.e., the criterion used to consider the trSNP) reflects the FDR for that given length.

## Polymorphism-to-Divergence Ratios

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 two-tail 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 SIV, Supplementary Material online, for details).

## 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 an  $MQ \geq 30$  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 nonsynonymous 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 1) the groups of individuals with genotype AA versus GG/GA and 2) the groups of individuals with genotype GG versus GA/AA. We computed differential expression for LAD1 for (1) and (2) using the DESeq package (Anders and Huber 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.

## Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

## Acknowledgments

The authors thank the sequencing group at the Max Planck Institute for Evolutionary Anthropology and Martin Kircher for data production. The authors also thank Felix M. Key, Gabriel Renaud, Jing Li, Kay Prüfer, and Joshua Schraiber for helpful discussions and suggestions, as well as four anonymous reviewers, whose comments improved substantially the manuscript. The authors are grateful to the Lola Ya

Bonobo sanctuary in Kinshasa, Democratic Republic Congo, and the Tchimpounga sanctuary, Jane Goodall Institute, Republic of Congo, for allowing access to the primate samples. This work was supported by the Max Planck Society. J.C.T. is supported by Fundação para a Ciência e a Tecnologia (FCT) within the Portuguese Ministry for Science and Education (SFRH/BD/77043/2011).

## References

- Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56–65.
- Aboobaker J, Wojnarowska FT, Bhogal B, Black MM. 1991. Chronic bullous dermatosis of childhood—clinical and immunological features seen in African patients. *Clin Exp Dermatol*. 16:160–164.
- Allison AC. 1956. The sickle-cell and haemoglobin C genes in some African populations. *Ann Hum Genet*. 21:67–89.
- Andrés AM. 2011. Balancing selection in the human genome. Encyclopedia of life sciences (eLS). Chichester: John Wiley & Sons Ltd.
- Andrés AM, Dennis MY, Kretschmar WW, Cannons JL, Lee-Lin SQ, Hurle B, Schwartzberg PL, Williamson SH, Bustamante CD, Nielsen R, et al. 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet*. 6:e1001157.
- Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, et al. 2009. Targets of balancing selection in the human genome. *Mol Biol Evol*. 26:2755–2764.
- Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol*. 11:R106.
- Asthana S, Schmidt S, Sunyaev S. 2005. A limited role for balancing selection. *Trends Genet*. 21:30–32.
- Bamshad MJ, Mummidhi S, Gonzalez E, Ahuja SS, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB, et al. 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. *Proc Natl Acad Sci U S A*. 99:10539–10544.
- Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*. 8:1499–1504.
- Cagliani R, Fumagalli M, Biasin M, Piacentini L, Riva S, Pozzoli U, Bonaglia MC, Bresolin N, Clerici M, Sironi M. 2010. Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene. *Hum Genet*. 128:577–588.
- Cagliani R, Guerini FR, Fumagalli M, Riva S, Agliardi C, Galimberti D, Pozzoli U, Goris A, Dubois B, Fenoglio C, 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, Nordborg M, Charlesworth D. 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.
- Clark AG. 1997. Neutral behavior of shared polymorphism. *Proc Natl Acad Sci U S A*. 94:7730–7734.
- Denguezli M, Ben Nejma B, Nouira R, Korbi S, Bardi R, Ayed K, Essoussi AS, Jomaa B. 1994. [Iga linear bullous dermatosis in children. A series of 12 Tunisian patients]. *Ann Dermatol Venereol*. 121:888–892.
- Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PLoS One* 7:e30377.
- Fan WM, Kasahara M, Gutknecht J, Klein D, Mayer WE, Jonker M, Klein J. 1989. Shared class II MHC polymorphisms between humans and chimpanzees. *Hum Immunol*. 26:107–121.
- Ferguson W, Dvora S, Fikes RW, Stone AC, Boissinot S. 2012. Long-term balancing selection at the antiviral gene OAS1 in Central African chimpanzees. *Mol Biol Evol*. 29:1093–1103.
- Ferrer-Admetlla A, Bosch E, Sikora M, Marqués-Bonet T, Ramírez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, et al. 2008. Balancing selection is the main force shaping the evolution of innate immunity genes. *J Immunol*. 181:1315–1322.
- Fischer A, Prüfer K, Good JM, Halbwax M, Wiebe V, Andre C, Atencia R, Mugisha L, Ptak SE, Paabo S. 2011. Bonobos fall within the genomic variation of chimpanzees. *PLoS One* 6:e21605.
- Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res*. 19: 199–212.
- Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet*. 7:e1002355.
- Gendzekhadze K, Norman PJ, Abi-Rached L, Graef T, Moesta AK, Layrisse Z, Parham P. 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 U S A*. 106: 18692–18697.
- Gigord LD, Macnair MR, Smithson A. 2001. Negative frequency-dependent selection maintains a dramatic flower color polymorphism in the rewardless orchid *Dactylorhiza sambucina* (L.) Soo. *Proc Natl Acad Sci U S A*. 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.
- Hodgkinson A, Eyre-Walker A. 2010. The genomic distribution and local context of coincident SNPs in human and chimpanzee. *Genome Biol Evol*. 2:547–557.
- Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet*. 12:756–766.
- Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. *PLoS Biol*. 7:e1000027.
- Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics* 120:831–840.
- Ishiko A, Shimizu H, Masunaga T, Hashimoto T, Dmochowski M, Wojnarowska F, Bhogal BS, Black MM, Nishikawa T. 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, Hellmann I. 2011. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. *Genome Biol Evol*. 3: 842–850.
- Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res*. 12: 656–664.
- Key FM, Teixeira JC, de Filippo C, Andrés AM. 2014. Advantageous diversity maintained by balancing selection in humans. *Curr Opin Genet Dev*. 29C:45–51.
- Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, et al. 2014. A draft map of the human proteome. *Nature* 509:575–581.
- Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res*. 40:e3.
- Kircher M, Stenzel U, Kelso J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol*. 10:R83.
- Klein J, Satta Y, O'Huigin C, Takahata N. 1993. The molecular descent of the major histocompatibility complex. *Annu Rev Immunol*. 11:269–295.
- Lappalainen T, Sammeth M, Friedlander MR, 't Hoen PA, Monlong J, Rivas MA, González-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501:506–511.

- Lawlor DA, Ward FE, Ennis PD, Jackson AP, Parham P. 1988. HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. *Nature* 335:268–271.
- Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339:1578–1582.
- Marinkovich MP, Taylor TB, Keene DR, Burgeson RE, Zone JJ. 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, Jonker M, Klein D, Ivanyi P, van Seventer G, Klein J. 1988. Nucleotide sequences of chimpanzee MHC class I alleles: evidence for trans-species mode of evolution. *EMBO J*. 7: 2765–2774.
- McKee PH, Calonje E, Granter SR, editors 2005. Pathology of the skin: with clinical correlations. Edinburgh: Philadelphia Elsevier Mosby.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20:1297–1303.
- Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*. 2010:pdb prot5448.
- Monia K, Aida K, Amel K, Ines Z, Becima F, Ridha KM. 2011. Linear IgA bullous dermatosis in tunisian children: 31 cases. *Indian J Dermatol*. 56:153–159.
- Motoki K, Megahed M, LaForgia S, Uitto J. 1997. Cloning and chromosomal mapping of mouse ladinin, a novel basement membrane zone component. *Genomics* 39:323–330.
- Muehlenbachs A, Fried M, Lachowitz J, Mutabingwa TK, Duffy PE. 2008. Natural selection of FLT1 alleles and their association with malaria resistance in utero. *Proc Natl Acad Sci U S A*. 105: 14488–14491.
- Muirhead CA, Glass NL, Slatkin M. 2002. Multilocus self-recognition systems in fungi as a cause of trans-species polymorphism. *Genetics* 161:633–641.
- Pasvol G, Weatherall DJ, Wilson RJ. 1978. Cellular mechanism for the protective effect of haemoglobin S against *P. falciparum* malaria. *Nature* 274:701–703.
- Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* 499: 471–475.
- Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486:527–531.
- Scheet P, Stephens M. 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, Mikkelsen AM, Hein J. 2001. Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. *Genetics* 159:1833–1844.
- Ségurel L, Thompson EE, Flutre T, Lovstad J, Venkat A, Margulis SW, Moysé J, Ross S, Gamble K, Sella G, et al. 2012. The ABO blood group is a trans-species polymorphism in primates. *Proc Natl Acad Sci U S A*. 109:18493–18498.
- Takahata N, Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 124:967–978.
- Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population-structure. *Evolution* 38:1358–1370.
- Wu C, Zhao K, Innan H, Nordborg M. 2004. The probability and chromosomal extent of trans-specific polymorphism. *Genetics* 168: 2363–2372.
- Wooding S, Kim UK, Bamshad MJ, Larsen J, Jorde LB, Drayna D. 2004. Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. *Am J Hum Genet*. 74:637–646.
- Wooding S, Stone AC, Dunn DM, Mummidi S, Jorde LB, Weiss RK, Ahuja S, Bamshad MJ. 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.