Oxytocin and vasotocin receptor variation sheds light into the evolution of human prosociality

Constantina Theofanopoulou[1,2,3,+], Alejandro Andirkó[2,3,+], Cedric Boeckx[2,3,4,*], Erich D. Jarvis[1,5,*]

[1] Laboratory of Neurogenetics of Language, Rockefeller University, New York, NY, USA
[2] Section of General Linguistics, Universitat de Barcelona
[3] Universitat de Barcelona Institute for Complex Systems
[4]ICREA
[5] Howard Hughes Medical Institute, Chevy Chase, MD, USA
[*]corresponding authors: cedric.boeckx@ub.edu, ejarvis@rockefeller.edu
[+]these authors contributed equally to this work

Abstract

Modern human lifestyle strongly depends on complex social traits like empathy, tolerance and cooperation. These diverse facets of social cognition have been associated with variation in the oxytocin receptor (*OTR*) and its sister genes, the vasotocin/vasopressin receptors (*VTR1A*/*AVPR1A* and *AVPR1B/VTR1B*). Here, we compared the full genomic sequences of these receptors between modern humans, archaic humans, and 12 non-human primate species, and identified sites that show heterozygous variation in modern humans and archaic humans distinct from variation in other primates, and that have associated literature. We performed variant clustering, pathogenicity prediction, regulation, linkage disequilibrium frequency and selection analyses on data in different modern-human populations. We found five sites with modern human specific variation, where the modern human allele is the major allele in the global population (*OTR*: rs1042778, rs237885, rs6770632; *VTR1A*: rs10877969; *VTR1B*: rs33985287). Among them, the *OTR*-rs6770632 was predicted to be the most functional. We found two sites where alleles (*OTR:* rs59190448 and rs237888)[1] present only in modern humans and archaic humans are under positive selection in modern humans, with rs237888 predicted to be a highly functional site. We identified three sites of convergent evolution between modern humans and bonobos (*OTR*: rs2228485 and rs237897; *VTR1A*: rs1042615), with *OTR*-rs2228485 ranking very highly in terms of functionality and being under balancing selection in modern humans. Our findings shed light on evolutionary questions of modern human and hominid prosociality, as well as on similarities in the social behavior between modern humans and bonobos.

Introduction

Modern humans are characterized by prosociality, a broad term that encompasses intraspecies empathy, social tolerance, cooperation and altruism. While our closest living relatives, the chimpanzees (*Pan Troglodytes*) and the bonobos (*Pan Paniscus*), live in highly organized social groups as well, present-day humans' social networks are larger and denser, powered by a complex social cognitive machinery[2]. Modern humans are also characterized by great intrasocial compassion and display a clear tendency to act in concert, to the extent that our species has been labeled as 'ultra-social'[3].

Less is known about the sociality of our extinct human relatives, the Neanderthals and the Denisovans. Paleogenomic studies show that both Neanderthal and Denisovan populations endured a population size bottleneck[4]. In addition to this, high inbreeding rates evident in both the Altai Neanderthal[5] and the Denisovan[6] genomes, as well as lower genetic diversity in Neanderthals compared to modern humans[7] suggest that their social groups were most likely much smaller than ours. According to Rogers and colleagues[8], Neanderthals were deeply subdivided into small isolated population groups with scarce contact between them, which may be associated with a social profile distinct from *Homo sapiens.*

A wide range of studies have highlighted a role of oxytocin and vasopressin/vasotocin receptor genes in social behaviors[9] and associated disorders, such as in Autism Spectrum Disorders (ASD)[10,11]. These include *OTR (*a.k.a. *OXTR), VTR1A (*a.k.a. *AVPR1A)* and *VTR1B (*a.k.a. *AVPR1B*), following the nomenclature proposed in[12]. These findings have led some researchers, most prominently Hare[13], to ascribe to these genes a key role in the emergence of human ultra-social behavior.

In this study, we searched for sequence variation in the *OTR-VTR1* family of genes (specifically, *OTR*, *VTR1A* and *VTR1B)* that could shed light on the evolution of prosociality in modern humans. Since fixed or nearly fixed changes have not been found so far in these genes in comparisons between MH and archaic humans (AH) or non-human primates (NHP) (chimpanzees and bonobos in particular)[5,14–16], we sought to identify potential polymorphic heterozygous sites in MH not found in AH or/and NHP, and we report those sites that have been associated with social features/deficits in the literature. We identified 5 sites with modern human specific variation, where the modern human alleles are the major alleles in the global population. Among them, sites in the *OTR* were in regulatory regions of the gene, active in the brain, and associated with prosocial behavior in modern humans. We also identified 3 convergent sites between modern humans and bonobos, a primate species that shows convergence of prosocial behaviors with humans, separate from chimpanzees. Some of the sites we report have been shown to be under balancing or positive selection in MH. Our findings shed light on the evolutionary questions on modern human and hominid prosociality, as well as on observations on similarities in the social behavior of MH and bonobos.

Results

Variation pattern analyses and associations with sociality

We performed alignments of the *OTR*, *VTR1A* and *VTR1B* genomic sequences (exon, introns, and surrounding regulatory regions 600 bp upstream and downstream) between seven high-coverage present-day human genomes (i.e. Modern Humans; MH), 14 Neanderthal genomes[5,17,18,19,20,21,22] and a Denisovan[6] genome (i.e. archaic humans; AH), and multiple genomes of 3 non-human primate (NHP) species (chimpanzee, bonobo, macaque; Supplementary Table 1). We also included several *VTR1A*-microsatellites that have been associated with social-related phenotypes[24,25]: RS3-$(CT)_4TT(CT)_8(GT)_{24}$; RS1-$(GATA)_{14}$; GT_{25}; and AVR-$(GT)_{14}(GA)_{13}(A)_8$. For the non-human primates, we additionally used available Single Nucleotide Variant (SNV)-data from bonobo, chimpanzee, and macaque populations (~30 to ~1000 individuals/species; **Supplementary Table 2**; Methods) to account for variation in non-human primates. We inferred the ancestral state of the sites we identified using Ortheus multialignments of 12 NHP species (Supplementary Note 1).

From the total number of Single Nucleotide Polymorphisms (SNPs) found on the *OTR* (5409 SNPs), *VTR1A* (2340 SNPs) and *VTR1B* (3031 SNPs) in the NCBI Variation Viewer (National Center for Biotechnology Information) in MH (GRCh38.p12 reference genome), only 79 had associated literature findings: *OTR* (55), *VTR1A* (7) and *VTR1B* (17). We excluded from our analysis those studies that review evidence from original studies, and we ended up with 61, consisting of 42 SNPs (285 studies) for *OTR*, 8 SNPs (24 studies) for *VTR1A*, and 11 SNPs (14 studies) for *VTR1B* (**Supplementary Tables 3, 4** and **5**). We classified these associations based on whether they were strongly associated to sociality (e.g. 'prosociality', 'social cognition'), possibly related to sociality (e.g. 'depression', 'Attention Deficit Hyperactivity Disorder'), or not related to sociality (e.g. 'diabetes'). We then calculated the percentage of the associations for each site, and then averaged all the percentages for each category relative to the total number of the SNPs.

MH variant sites in *OTR* had the highest association with sociality (**Supplementary Table 3**): 72% of the findings were sociality-related (e.g. 'empathy', 'social temperament', 'face recognition'); 14% possibly related to sociality (e.g. 'obsessive compulsion disorder', depression); while the remaining 14% was not related to sociality (e.g. 'bulimia', 'overeating', 'diabetes'). These differences were significant (Chi-squared test; $P < 0.0001$). In contrast, variants in *VTR1A* had a weaker association, with 27% being related to sociality (e.g. 'moral judgement', 'aggression', 'autism'), 6% possibly related to sociality (e.g. 'susceptibility for panic disorder'), and 68% to other phenotypes (e.g. 'nicotine dependence', 'heroin addiction', 'hypertension'; **Supplementary Table 4**; Chi-squared test; $P = 0.0049$). Concerning *VTR1B*, 50% of the findings were related to sociality (e.g. 'emotional empathy', 'aggression'), 32% possibly related to sociality ('bipolar disorder', 'psychotic features'), and 17% were not (e.g. 'elevated body mass index'; **Supplementary Table 5**). Unlike the other two genes, these values for *VTR1B* were not significantly different between groups (Chi-squared test; $P = 0.3417$). Lastly, the difference between the proportions of the studies related to sociality were significant for *OTR* versus *VTR1A* (Chi-squared test; $P < 0.0001$), but not significant for *OTR* versus *VTR1B* (Chi-squared test; $P = 0.0771$) and *VTR1A* versus *VTR1B* (Chi-squared test; $P = 0.1836$; **Supplementary Table 6**). We interpret these findings as evidence for a higher functional association to sociality for *OTR* over

VTR1A and *VTR1B*, although we are aware it could be a result of publication bias favoring sociality studies with *OTR*.

In total, of the 61 sites analyzed, 29 had associations with sociality (**Supplementary Tables 7-9**), while simultaneously showed a variation pattern found only in MH, in AH and MH, or in MH and a NHP species: 19 *OTR*-SNPs, 7 *VTR1A*-SNPs and 3 *VTR1B*-SNPs. Based on their variation patterns in MH, AH and NHP, we classified the 29 sites into five categories (**Figure 1**): 1) Modern Human Unique (MHU) for sites where the heterozyogous allelic variation is unique to MH (e.g. T/A in MH and A/A in all other species); 2) Modern Human Specific (MHS) for sites where MH have a specific variation pattern (e.g. T/A) different from that in other species (e.g. C/A); 3) Homo Unique (HU) for sites that are variant in both MH and AH in the same fashion (e.g. T/A), while the rest of the NHP are invariant (e.g. A/A) or have a different variant (e.g. C/A); 4) Homo Specific (HS) for sites that are variant in MH (e.g. T/A), and invariant in AH (e.g. T/T), different from all NHP species (e.g. A/A); And 5) MH-NHP for sites with convergent variation between MH (e.g. T/A) and another NHP (e.g. T/A in bonobo), where other species/populations are invariant for the ancestral allele (e.g. A/A). Of these 29 sites associated with sociality, most (13) were MHU in *OTR*, while the 3 in *VTR1B* were MHS. The others were in a mixture of genes, with 4 HS sites in *OTR* and 1 *VTR1A*; 2 HU sites in *OTR* and 1 in *VTR1A*; and 3 convergent sites between MH and bonobos in *OTR* and 1 in *VTR1A* (**Figure 1, Table 1**).

Ancestral allele and functional association analyses revealed that in almost all cases, the ancestral and the MHU, MHS, HU or HS alleles showed mixed associations with sociality, i.e., the same allele was associated with both 'prosocial' and 'antisocial' features in the literature. In other words, we did not find a general trend where the ancestral allele would be associated with 'antisocial' features, and the new allele with 'prosocial' ones, or the other way around (**Supplementary Tables 7-9**). For example, the MHU *OTR*-rs1042778(G), which is highly present in MH populations, has been associated with ASD[26–28], aggression in males, but also with prosocial fund allocations[29] and high scores in altruistic and comforting behaviors[30] (**Supplementary Table 7**). The ancestral T allele has been associated with diminished parental care[31] and panic/aggressive behaviors[28] on the one hand, but also more likely recovery from low maternal emotional warmth on the other[32]. A similar trend of mixed social associations for each variant, or absence of evidence for one of the allelic variants, was found for the rest of the sites (**Supplementary Table 7-9**).

Pathogenicity prediction analyses

Most of these sites were located in regulatory regions of the genes, but some in exons giving rise to missense (*OTR*-rs4686302 [HS], *VTR1B*-rs28632197 [MHS] and *VTR1A*-rs1042615 [MH-NHP convergence]) or synonymous mutations (*VTR1A*-rs2228485 [MH-NHP convergence] and *VTR1B*-rs28676508 [MHS]) (**Table 1**). We performed pathogenicity prediction analyses using genome wide scores of conservation that predict deleteriousness of mutations (CADD[33] [Combined Annotation Dependent Depletion] and GERP[34] [Genomic Evolutionary Rate Profiling]), and algorithms that evaluate the impact of missense variants (Mutation Assessor[35], REVEL[36] [Rare Exome Variant Ensemble Learner], MetaLR[37] [Meta Logistic Regression], SIFT[38] [Sorting Tolerant From Intolerant] and PolyPhen-2[39]) (**Table 1**). CADD scores that predict variant deleteriousness showed that MH-NHP *OTR*-rs2228485(A) and *VTR1A*-rs1042615(G, C, T) alleles are likely deleterious (i.e. scores >20 predict deleteriousness), while HS *OTR*-rs4686302(T) also scored very close to this value (19; **Table 1**). Scores for most other sites ranged from ~0 to 6. GERP tests showed positive evolutionary rate scores (as defined in Methods) only for MHU *OTR*-rs2254298 and HS *OTR*-rs13316193 (**Table 1**), which represent highly conserved sites, that are likely to be functional. Combined REVEL, MetaLR, SIFT and PolyPhen-2 scores classified HS *OTR*-rs4686302 as more likely benign, with only the Mutation Assessor classifying it as likely deleterious (**Table 1**). MH-NHP *VTR1A*-rs1042615 was ranked by most of the aforementioned analyses as deleterious or likely deleterious, with only MetaLR ranking it as benign. SIFT and PolyPhen-2 tests categorized MHS *VTR1B*-rs28632197 as benign. These findings suggest that despite their deleterious potential, these sites have been maintained in the modern human population but not in other primates, perhaps due to selection differences in social behavior.

Regulation analyses

We performed regulation analyses using RegulomeDB[40] and the Ensembl regulation data resources[41]. RegulomeDB make use of large datasets, including ChIP-seq, FAIRE-seq and DNase I hypersensitive information for a variety of important regulatory factors across a diverse set of cell types; this includes

chromatin state information for over 100 cell types, transcription factor binding sites (including Position-Weight Matrix and DNase Footprinting), and expression quantitative trait loci (eQTL) information allowing the association of distal sites with gene promoters. The Ensembl resources provide a functional annotation of the regulatory elements in the human genome, including data on epigenetic marks, transcription factor binding and DNA methylation, as well as microarray probe mappings.

Our RegulomeDB analyses ranked highly the MHU *OTR*-rs237889 variant site as lying in a functional genomic location (rank 1f; **Table 2**). Consistent with these findings, this same site is found in a genomic location with peak Chip-seq and DNase-seq signal for open chromatin in the brain and other cell types (**Figure 2**). This site is also located in an eQTL affecting expression of *CAMK1* and maps to the binding sites of the *EGR2* and *ZFHX2* transcription factors (**Table 2**). The second most highly ranked site was the MH-NHP *OTR*-rs2228485 allele (rank 2b). ChIP-seq, DNase-seq and FAIRE-seq data indicate that this site is in a genomic region active in the brain, and in particular the dopaminergic substantia nigra neurons and one of its projection targets, the caudate nucleus of the striatum, as well as the hippocampus, other parts of the temporal lobe, the cingulate and angular gyri, and the middle frontal area 46 (**Table 2**). This site is implicated in binding with *CTCF*, *EZH2*, *POLR2A*, *RAD21* and *IKZF1* transcription factors, but with the SNP potentially altering the binding for *PAX6* and *ZSCAN4C* (**Table 2**).

MH-NHP *OTR*-rs2228485, MHU *OTR*-rs2268493, HU *OTR*-rs60902022, HU *OTR*-rs237888 and MHS *VTR1B*-rs28632197 were all were less likely to affect regulation (rank 3a), but still had open chromatin regions that either fall within the binding site of different proteins, or alter gene motifs, or both, and showed activity in brain regions like the cingulate gyrus and the hippocampus (**Table 2; Figures 2a, 2b**). The remaining sites (*OTR*: rs2268490, rs4686302, rs4686302, rs237917, rs6770632, rs59190448 and rs9872310; *VTR1A*: rs10877969 and rs1042615; *VTR1B*: rs33985287) did not rank highly in changing regulation according to the RegulomeDB scores (ranks 4-6), but were also found in open chromatin or promoter regions (according to Ensembl or RegulomeDBdata), or in a transcription factor binding domain (**Figure 2a**), and all showed activity in the brain. A total of 8 of these 15 sites are located within the *CTCF* and the *POLR2A* transcription factor binding motifs (**Table 2; Supplementary Table 10**). These findings indicate that variant sites associated with sociality would cause variant brain gene expression among the MH population.

Linkage disequilibrium analysis

We ran Linkage Disequilibrium analysis (using LDmatrix[42]) to test if any of the sites we identified are co-inherited, using data from all populations included in the 1000 Genomes Project and a cut off of $R^2 > 0.8$. This analysis revealed four haplotype blocks: one containing co-inheritances among rs60902022, rs13316193 and rs11131149 in *OTR*; a second between rs10784339 and rs10877969 in *VTR1A*; a third between rs11174811 and rs3021529 also in *VTR1A*; and a fourth between rs33985287 and rs28676508 in *VTR1B* (**Figure 3, Supplementary Table 11**). This suggests that the combinations of these variant alleles might act in concert to shape functional differences that end up affecting phenotypes controlled by these genes.

Population data and selection analysis

We further analyzed the human allele frequency data for these sites (1000 Genomes Project[43] Phase 3 and gnomAD[44] v2.1.1 and v3) and the literature for selection on these sites (**Table 1, Supplementary Table 12, Extended Data Table 1**). We used the supplementary information in Schaschl et al. 2015[1], and filtered the results of their FDIST, Bayescan and extended Lewontin and Krakauer (FLK) tests, reporting here only those on our identified SNPs with significant values (p or q values < 0.05) (**Extended Data Table 1**). For bonobos, we also analyzed the data from 13 bonobos by Kovalaskas and colleagues[45], where strong positive selection signals on *OTR* and *VTR1A* were identified.

We found four MHU and one MHS case where the new allele is currently the major allele in MH populations: (MHU: *OTR*-rs6770632(C), rs1042778(G), rs237885(T); MHS: *VTR1A*-rs10877969(T); *VTR1B*-rs33985287(T); **Table 1**). Several of these sites have a high derived allele frequency of ≥70% globally (*OTR*-rs6770632(C), *VTR1A*-rs10877969(T), *VTR1B*-rs33985287(T)), while *OTR*-rs1042778(G) and *OTR*-rs237885(T) showed lower frequencies (59% and 51%, respectively). Similarly, we found one HU and one HS case, where the new hominin (AH or MH) allele is the major allele in the global MH population (HU: *OTR*-rs59190448(G); HS: *OTR*-rs237888(T); **Table 1**). Both sites show derived allele frequencies of >85%,

and they show signs of positive selection. Specifically, based on our filtering of the results in[1], all three approaches (FDIST, BayeScan and FLK) identified the *OTR*-rs59190448(G) to be strongly under positive directional selection (FDIST p=0.0004, BayeScan q=0.0004, F.LK p=0.001), while only the FLK test (p=0.011) detected the *OTR*-rs237888 to be under positive selection (**Extended Data Table 1**). Interestingly, both the *OTR*-rs237888(T) and the *OTR*-rs59190448(G) are present in the majority of the AH genomes (6/8 for *OTR*-rs237888(T) and 5/7 for *OTR*-rs59190448(G); **Supplementary Table 1**); although, we acknowledge that the genomes of a small number of AH individuals currently available might not be representative of the general archaic population. These results, though tentative, show a similar frequency trend in AH and MH.

Several sites where we encountered a new allele either in AH or in MH are currently the minor allele in MH populations, with some of them being present at low percentages in the global population (<25%) (*OTR*-rs9872310, rs2268493, rs2254298, rs4686302, rs237911; *VTR1A*-rs11174811, rs3803107, rs3021529, rs3759292; and *VTR1B*-rs28676508, rs28632197). Our filtering of the results in[1] showed that some of them are under balancing selection: *OTR*-rs237911 (FDIST, p= 0.0485), *VTR1A*-rs11174811 (FDIST, p=0; BayeScan, q=0.0336) and *VTR1A*-rs3021529 (FDIST, p= 0.0001; BayeScan q= 0.0266); only the *VTR1A*-rs3759292 was found to be under positive selection (F.LK, p=0.001; **Extended Data Table 1**). Interestingly, *VTR1A*-rs3759292 is also in a region that is found deleted in 0.3% of the macaques sampled on that site (739; **Supplementary Table 2**), and according to Donaldson and colleagues[25], who found sites with both variation and deletion events in the MH and NHP *VTR1A* microsatellite, such regions might have influenced sociobehavioral traits during primate evolution. MH and bonobos showed convergent variation on three other sites (*OTR*-rs2228485 and rs237897; *VTR1A*-rs1042615). The *OTR*-rs2228485 allele was present in all the 13 bonobos of one population[46] (87% of our total sample of 15 bonobos), but the alternate allele was present in the two reference genomes (13%). Although in all these sites, the convergent allele was not the major allele in the global MH population, in some MH populations it is present at ~50% (**Table 1, Supplementary Table 12**). In particular, the convergent *OTR*-rs2228485 G-allele that was found at 87% among the bonobo population, at 28% in the global MH population, while at 46% in the American population. *OTR*-rs237897(A) was found in 40% of the MH global population (50% in Europeans) and 13% in bonobos. *VTR1A*-rs1042615(A) reached 35% in the global population (49% in Europeans), and 87% in bonobos. Of these sites, only the *OTR*-rs2228485 was found to show signs of balancing selection (FDIST, p=0.0149) in MH. None of these convergent sites were among the specific sites of strong positive selection identified in bonobos[45], but several of these sites were located in close vicinity to our identified sites in the *OTR* and *VTR1A* (e.g. one site was found at a ~14kb distance from *VTR1A*-rs1042615). These findings suggest a correlation of several sites associated with sociality also showing convergent changes in modern humans and bonobos, not found in chimpanzees and other non-human primates examined.

Discussion

In this study, we find multiple cases of allelic heterozygosity variation in the *OTR*, *VTR1A* and *VTR1B* genes that is specific to MH or AH and MH, where one allele is unique to them among primates and the other shared with other primates. Variants in all three genes have been associated with sociality, although more so for *OTR* than *VTR1A* and *VTR1B*. As it is known that heterozygous variants can have different impacts on genes and traits than their homozygous counterparts, such as the protective affect for sickle cell anemia[47], we suggest that these variants could have a synergistic heterozygous impact on the evolution of social behavior in hominids. However, because of the population variation, we believe that it may not be a specific allele that is important for increased sociality, but the combination of heterozygous alleles among the population that may influence the interactions differently than a population that were purely homozygous. SNPs can act as general risk factors, and not as specific modifier alleles in certain behavioral or biological contexts (i.e. disorder specific effects)[48]. Certain alleles associated with more social deficits in one disorder (e.g. ASD) had either no effect, or conferred a benefit with respect to social abilities in another disorder (e.g. ADHD). Additionally, the impact of the alleles of a certain variant site is different depending on environmental factors, because environmental processes may be differentially accentuated through hormone signaling. We thus focus our interpretations on the functional and frequency analyses.

We consider that of the MHU and MHS SNPs, only those five where the new allele is the major allele in MH populations could be responsible for beneficial changes in MH relevant to sociality (*OTR*-rs6770632, rs1042778, rs237885; *VTR1A*-rs10877969; *VTR1B*-rs33985287), especially those found at ≥70% (*OTR*-

rs6770632, *VTR1A*-rs10877969, *VTR1B*-rs33985287). *OTR*-rs6770632 is the only site among these three that falls in open chromatin regions active in the brain, and particularly in the cingulate gyrus, a brain region involved in social cognition[49]. Moreover, this site is in a genomic region containing the binding domain of *EZH2*, a gene that is involved in the etiology of memory impairments[50], autism[51] and Weaver syndrome, an overgrowth/intellectual disability syndrome[52].

These MHS sites can shed light on several working hypotheses that account for MH prosociality. Among them, the 'self-domestication hypothesis'[13,53–56] posits that certain physiological and behavioral traits that MH share with domesticated animals support a significant turning point exclusive of MH on the prosocial continuum. Interestingly, Hare[13] and Theofanopoulou[54] have hypothesized that human self-domestication was likely facilitated by changes in the oxytocin/vasotocin system, based in part on studies showing that oxytocin and vasotocin receptors are under relaxed selective constraint in domesticated species[57], and their gene expression and methylation patterns differ in wild versus domesticated species[58,59]. We also found that some of the variants found only in MH among primates (*VTR1A*-rs11174811 and *VTR1A*-rs3021529) show signs of relaxed selection.

Linkage disequilibrium analyses reveal more intricate relationships between alleles that might act in unison, likely conferring differential effects in specific MH populations. For example, the MHU *VTR1A*-rs10877969 variant co-inherited with the HS *VTR1A*-rs10784339, although not present in high percentages globally (31%), could be responsible for differences in the African population (58% in the African population). Similarly, other alleles that are found in low percentages in MH but showed highly likely functional results, like the *OTR*-rs237889 (30% in the global population; highest score (1f) in RegulomeDB analysis), might be responsible for characteristics of the Japanese population, where the MHS allele is more frequent (48%). According to Schaschl and colleagues[1] who studied selection patterns of *OTR* and *VTR1A* SNPs in different human populations, the different population distribution of these alleles could influence the emergence of differences in socio-cultural practices on interpersonal proximity and touch, emotional suppression, social or emotional support, use of facial expression, inter-individual distance, and even susceptibility to ASD.

HU sites, where AH is homozygous with a new mutation not found in NHP, and MH are variant, can offer information on those SNPs that were likely present in AH and were selected for or purged out in the MH populations. *OTR*-rs59190448(G) and *OTR*-rs237888 are the only two sites where the HU allele ended up being positively selected in MH[1], possibly resulting in beneficial phenotypes. With the open chromatin region for *OTR*-rs237888 also in the cingulate gyrus, we propose that this site along with the aforementioned *OTR*-rs59190448, could account for a more prosocial phenotype in the human lineage compared to our non-human primate ancestors at the time of the *Pan-Homo* lineage split[60,61]. Contrariwise, some of the other HU alleles that were found in open chromatin or conserved sites, have not been highly retained in the global MH population (*OTR*: rs9872310 [open chromatin], rs13316193 [conserved site], rs4686302 [open chromatin, mixed results on whether it is benign or deleterious]; *VTR1A:* rs3803107).

For the three sites of convergent evolution between MH and bonobos, *OTR*-rs2228485 was found to be under balancing selection in MH and is also one of the most highly ranked sites in our RegulomeDB analysis, pointing to putative high functionality. Functionality in this site with balancing selection might also contribute to phenotypic plasticity, as has been shown for other sites[62]. Only the *OTR*-rs2228485(G) and the *VTR1A*-rs1042615(A) were the major alleles in our bonobo sample, but in order to conclusively resolve the state of these sites in bonobos, larger population samples will be needed. Nonetheless, our findings on convergent changes for MH with bonobos could be insightful for understanding the posited similarities in prosociality between MH and bonobos, and the differences of both compared to chimpanzees[63,64]. Added to that, according to a recent study[45], bonobos, since their split from chimpanzees, show strong positive selection signals on variants located in genes of the oxytocin and vasotocin pathway (including *OTR* and *VTR1A*), something they consider as evidence for reduced aggression and self-domestication in bonobos.

In conclusion, we found more evidence (5 sites in oxytocin and vasotocin receptors) pointing towards a prosocial shift in MH, and less (2 sites) supporting a similar shift in the hominin populations altogether.

Although the fact that there are vastly more MH genomes currently available that might tip the balance towards MH specificity, the contrasting patterns obtained for oxytocin and vasotocin receptors suggest that our results cannot be fully reduced to the number of genomes available. Other receptors and neural ligands besides *OT* and *VT* with early changes in hominids, such as dopamine[61] and β-endorphines[60], are possible. When considering these hypotheses, one will need to consider that all of these ligands are known to interact[65], so it could be that the early changes in *OTR* identified here formed part of a broader set of changes that set the stage for our prosocial profile. Our results and interpretations are also broadly compatible with information based on the fossil record, paleogenomic evidence[5–8], and with behavioral differences between chimpanzees and bonobos[63,64].

Methods

Variation pattern analyses

We retrieved the *OTR*, *VTR1A* and *VTR1B* DNA sequences from the following sources: the publicly available genomes of 14 Neanderthal individuals[5,17,18,19,20,21,22] and a Denisovan[6]; seven high-coverage present-day human genomes (San(HGDP01036), Mbuti(HGDP00982), Karitiana(HGDP01015), Yoruba(HGDP00936), Dinka(DNK07), French(HGDP00533) and Han(HGDP00775), originally sequenced for[5]; 1000 Genomes project data Phase 3[43] (as shown in Ensembl[66]); a chimpanzee (*Pan Troglodytes*) (CHIMP2.1.4/panTro4), bonobo (*Pan Paniscus*) (panpan1.1/panPan2) and rhesus macaque (*Macaca Mulatta*) (BMC Mmul_8.0.1/rheMac8).

We ran a multi-alignment between the modern human, archaic human, chimpanzee, bonobo and macaque gene sequences of *OTR*, *VTR1A* and *VTR1B*. Of the differences we found, we focused on those which are polymorphic in MH with literature associated with sociality or social deficits. We performed the alignments using the 'Phylogenetic Context' tool built in Ensembl[66], tools in the Max Planck for Evolutionary Anthropology Ancient Genome Browser (https://bioinf.eva.mpg.de/jbrowse/), including MUSCLE[67] and MView[68] to generate the alignments, and Aliview[69], Decipher for R[70], Bedtools to edit and analyze them. We used all the genomic sequence of the genes we aligned, as provided in the standard layout of the files of the genomic sequences in the Ensembl database, namely with 600 bp upstream and downstream. We defined the genomic sequences in the same way when we extracted the gene sequences from the archaic genomes. The only additional regions we included ad hoc in our analysis that fall further out of 600bp window were those of several *VTR1A*-microsatellites that have been associated with social-related phenotypes[24,25] (RS3-$(CT)_4TT(CT)_8(GT)_{24}$, RS1-$(GATA)_{14}$, GT_{25}, and AVR-$(GT)_{14}(GA)_{13}(A)_8$).

We then aligned these sites with the rest of the available Neanderthal genomes (AH: Spy[18] (2 individuals), Goyet[19], Les Cottés[20] (5 individuals), Hohlenstein-Stadel[21], Scladina[21], Chagyrskaya[23]) and additional NHP genomes, chimpanzee (*Pan Troglodytes*) (Pan_tro 3.0/ panTro5 and Clint_PTRv2/ panTro6), bonobo (*Pan Paniscus*) (Max-Planck Institute panpan1/panPan1) and Rhesus macaque (*Macaca Mulatta*) (Mmul_10/ rheMac10). Of the AH sites, we included only sequence reads that had a mapping quality of ≥25. Due to low-quality DNA for most of the ancient samples, and given the heterogeneity of procedures for data production in AH genome sequencing, we considered variant alleles only those present in at least two AH genomes, to avoid false positives due to sequencing errors. Following[71], we filtered out the samples that strongly deviated from the remaining individuals in the dataset. This filtering of samples only affected two major alleles discussed in the study: the MH-bonobo convergent *VTR1A*-rs1042615(A), that is also found in the Denisovan sequence; and the MHS *OTR*-rs6770632(C) allele that is also found in the Altai Neanderthal sequence. All alleles we retrieved can be found in Supplementary Table 1.

To assess variation in NHP, we used Single Nucleotide Variant (SNV)-data from: bonobos (13 bonobos from[46], 27 individuals with only *OTR* and *VTR1A* data from[72], 113 individuals with only *VTR1A* data from[73]); chimpanzees (25 individuals from[46], 35[72] and 62[74] individuals with only *VTR1A* and *OTR* data); and macaques (1,234 individuals from the mGAP database[75] and dbSNP 150 in Ensembl[66], where the mGAP data have been incorporated) (Supplementary Table 2). The SNV-data from bonobos and chimpanzees from[46] were lifted over to the GRCh38/hg38 version of the human genome (*Homo sapiens*), while the macaque SNVs[75] were lifted to the GRCh37/hg19 version of the human genome; they were lifted from different human references because

these are where the different primate data sets were mapped to. We lastly used information from several other publicly available genomes of these species, since they were sequenced from different individuals (chimpanzee: Pan_tro 3.0/ panTro5 and Clint_PTRv2/ panTro6, bonobo: Max-Planck Institute panpan1/panPan1, and rhesus macaque: Mmul_10/ rheMac10).

Ancestral allele inference

In order to infer the ancestral alleles of the identified sites, we aligned the MH sequences of *OTR*, *VTR1A* and *VTR1B* against 12 non-human primate species, and inferred the ancestral sequences through Ortheus[76] using the 'Phylogenetic Context' tool built in Ensembl[66]. Ortheus is a probabilistic method that infers the ancestral allele, using a phylogenetic model, that incorporate gaps, to infer insertion and deletion events. Ancestral sequences are predicted for each node of the phylogenetic tree. Multiple alignments of all the sites we identified in 12 primate species, with the inferred ancestral sequences in each node, can be found in Supplementary Note 1. Although we found several other putative sites of convergent evolution between humans and other non-human primates (e.g. mouse lemur, olive baboon, gelada etc.; **Supplementary Note 1**), in absence of variation data in these species we can't determine the reliability of this convergence as we have done with bonobo population data.

Association studies analysis

We went through all the association studies identified in the NCBI Variation Viewer[77] as of October 2020 (e.g. for *OTR*: https://www.ncbi.nlm.nih.gov/variation/view/?q=OXTR) for SNPs that have publications (Filters: 'Variant type' (: Single Nucleotide Polymorphism), and 'Has publications' (: Yes)), and report the association(s) identified for each SNP site (**Supplementary Tables 3, 4 and 5**). We then classified these associations on whether they were strongly related to sociality (e.g. 'prosociality', 'social cognition'), 'possibly related to sociality' (e.g. 'depression', 'Attention Deficit Hyperactivity Disorder'), or not (e.g. 'diabetes'). We calculated the percentage of the associations for each site first, to then add all the percentages for each category together averaged out by the total number of the SNPs. Although there are more studies on the *OTR* SNPs in total, mostly due to replication bias, we normalized this factor, by calculating an association-percentage for each SNP first (whether it came from one or several studies), and then used the number of the SNPs to average out the final results (instead of the number of studies). We calculated p values between the percentages, using a Chi-squared test (n exact sample sizes, degrees of freedom and confidence intervals are noted in **Supplementary Tables 3, 4 and 5**). We lastly calculated the difference between the proportions of the studies related to sociality for each pair-wise comparison of the analyzed genes, using a Chi-squared test (n exact sample sizes, degrees of freedom and confidence intervals are noted in **Supplementary Table 6**). All the sites we report have studies that we classified as clearly related to sociality, with the exception of rs10784339 that has been associated with substance use disorders (**Supplementary Tables 7-9**). Although we classified those studies as not related to sociality, we included this site in the analysis, recognizing that there might be subjectivity in the classification of substance use disorders as being related to sociality or not.

For some sites where we identified variation patterns we wanted to study further in terms of disorders, we performed an additional exhaustive literature research in the National Center for Biotechnology Information[77] (https://www.ncbi.nlm.nih.gov/pubmed/), SNPedia (http://snpedia.com) and Google Scholar (https://scholar.google.com/platforms) (as of October 2020) for the clinical significance of each one of the SNPs we identified in present-day human populations. On Supplementary Tables 10-12, we have included most of studies that report an association of these variants, a short description of their exact impact, as well as the trial sample of each study. We have not included review studies that recapitulate findings of original studies or studies showing negative results.

Pathogenicity prediction analyses

We performed pathogenicity prediction analyses using the following seven tools:

1) CADD[33] (Combined Annotation Dependent Depletion), which scores the predicted deleteriousness of SNVs and insertion/deletions variants in the human genome by integrating multiple annotations including conservation and functional information into one metric. In our analysis, we display scores above 20 as likely deleterious and scores below 20 as likely benign.

2) GERP[34] (Genomic Evolutionary Rate Profiling), which identifies constrained loci in multiple sequence alignments by comparing the level of substitution observed to that expected if there was no functional

constraint. Positive scores represent highly-conserved positions while negative scores represent highly-variable positions.

3) Mutation Assessor[35], which predicts the functional impact of amino-acid substitutions in proteins using the evolutionary conservation of the affected amino acid in protein homologs. The rank score is between 0 and 1, with variants with higher scores being more likely to be deleterious.

4) REVEL[36] (Rare Exome Variant Ensemble Learner), which is an ensemble method for predicting the pathogenicity of missense variants. It integrates scores from MutPred, FATHMM v2.3, VEST 3.0, PolyPhen-2, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP++, SiPhy, phyloP, and phastCons. Scores range from 0 to 1, and variants with higher scores are predicted to be more likely to be pathogenic. We considered scores above 0.5 as 'likely disease causing' and below 0.5 is 'likely benign'.

5) MetaLR[37] (Meta Logistic Regression), which uses logistic regression to integrate nine independent variant deleteriousness scores and allele frequency information to predict the deleteriousness of missense variants. Variants are classified as 'tolerated' or 'damaging'; scores rank between 0 and 1, and variants with higher scores are more likely to be deleterious.

6) SIFT[38] (Sorting Tolerant From Intolerant), which predicts whether an amino acid substitution is likely to affect protein function based on sequence homology and the physico-chemical similarity between the alternate amino acids. Scores < 0.05 are classified as 'deleterious' and all others as 'tolerated'.

7) PolyPhen-2[39], which predicts the effect of an amino acid substitution on the structure and function of a protein using sequence homology, Pfam annotations, 3D structures from PDB where available, and a number of other databases and tools (including DSSP, ncoils etc.). The PolyPhen-2 score represents the probability that a substitution is damaging, so values nearer 1 are more confidently predicted to be deleterious (note that this is opposite to SIFT).

Regulation analyses

We performed regulation analyses using RegulomeDB[40] and Ensembl regulation data resources[41]. RegulomeDB analysis make use of large datasets, including ChIP-seq, FAIRE-seq and DNase I hypersensitive information for a variety of important regulatory factors across a diverse set of cell types, chromatin state information across over 100 cell types, binding sites of transcription factors (including Position-Weight Matrix and DNase Footprinting), and expression quantitative trait loci (eQTL) information allowing the association of distal sites with gene promoters. The Ensembl resources provide a functional annotation of the regulatory elements in the human genome, including data on epigenetic marks, transcription factor binding and DNA methylation, as well as microarray probe mappings (**Table 2, Figure 2, Supplementary Table 10**).

RegulomeDB has developed a heuristic scoring system based on functional confidence of a variant. Category 1 variants are those that are known eQTLs for genes, and thus have been shown to be associated with expression. Within Category 1, subcategories indicate additional annotations from the most confident (1a, which has transcription factor (TF) binding, a motif for that TF, and a DNase footprint) to the least confident (1f, which has only TF binding or a DNase peak). According to Boyle and colleagues[40] who developed the system: '*the additional categories represent analogous annotations to Category 1 but without eQTL data and, thus, no known direct effect on binding. Category 2(a–c) demonstrates direct evidence of binding through ChIP-seq and DNase with either a matched PWM (positional weight matrix) to the ChIP-seq factor or a DNase footprint. Category 3(a–b) is considered less confident in affecting binding due to a more incomplete set of evidence. These sites have ChIP-seq evidence and either a motif that matches the ChIP-seq data but no DNase evidence, or DNase evidence and any other motif. Finally, Categories 4–6 lack evidence of the variant actually disrupting the site of binding. These include DNase and ChIP-seq evidence (Category 4), DNase or ChIP-seq evidence (Category 5), or any single annotation not in the above categories (Category 6).*' A list of the scores of all variants studied can be found in Supplementary Table 7.

Linkage Disequilibrium analysis

We ran Linkage Disequilibrium analysis (using LDmatrix[42]) to test if any of the sites we identified get coinherited, using data from all population included in the 1000 Genomes Project and a cut off of $R^2 > 0.8$. LDMatrix (https://ldlink.nci.nih.gov/?tab=ldmatrix) creates an interactive heatmap matrix of pairwise linkage disequilibrium statistics. R^2 and D' scores for all sites' combinations can be found in Supplementary Table 8.

Population data and selection analysis

We collected human allele frequency data for our identified sites using data from the 1000 Genomes Project[43] (Phase 3) and gnomAD v2.1.1 and v3[44]. We listed the minor allele frequency (MAF) and the highest MAF in present day human populations in Table 1, and specific frequencies only from the 1000 Genomes Project in Supplementary Table 9. For bonobos, we also analyzed the data from 13 bonobos by Kovalaskas and colleagues[45] (raw data provided by Sarah Kovalaskas), where strong positive selection signals on *OTR* and *VTR1A* were identified.

We also reviewed the literature for studies on selection signals of these sites in MH. We used the supplementary information found in Schaschl et al. 2015[1], and filtered the results of their FDIST, Bayescan and extended Lewontin and Krakauer (FLK) tests, reporting only those on our identified SNPs with significant values (p or q values < 0.05). In their study, Schaschl et al. 2015[1] reported in the main text only those SNPs with p or q values < 0.005. Schaschl et al. 2015[1] analyzed 14 human populations (n = 1092) from the 1000 Genomes database and aimed to identify SNPs under selection employing two tests that use Bayesian models to detect F_{ST} outliers (FDIST and Bayescan) and the FLK test that compares patterns of differences between allele frequencies in several populations relative to their expectation under neutral evolution (details for each analysis in the Methods of Schaschl et al. 2015[1]).

Figures

a

Variant ID	Gene	MH	AH	Chimpanzee	Bonobo	Macaque	Type
rs9872310	OTR	A/G	G/G	A/A	A/A	A/A	HS
rs6770632	OTR	C/A	A/A	A/A	A/A	A/A	MHU
rs1042778	OTR	G/T	T/T	T/T	T/T	T/T	MHU
rs237885	OTR	T/G	G/G	G/G	G/G	G/G	MHU
rs2268490	OTR	C/T	C/C	C/C	C/C	C/C	MHU
rs237888	OTR	T/C	T/C	C/C	C/C	C/C	HU
rs2268493	OTR	T/C	T/T	T/T	T/T	T/T	MHU
rs2254298	OTR	G/A	G/G	G/G	G/G	G/G	MHU
rs60902022	OTR	T/C	T/C	T/T	T/T	T/T	HU
rs237889	OTR	T/C	C/C	C/C	C/C	C/C	MHU
rs13316193	OTR	T/C	C/C	T/T	T/T	T/T	HS
rs11131149	OTR	G/A	A/A	G/G	G/G	A/A	-
rs59190448	OTR	G/A	G/G	A/A	A/A	A/A	HS
rs53576	OTR	A/G	G/G	G/G	G/G	G/G	MHU
rs237897	OTR	A/G	G/G	G/G	A/G	G/G	MH-NHP
rs4686302	OTR	C/T	T/T	C/C	C/C	C/C	HS
rs2228485	OTR	G/A	A/A	A/A	G/A	A/A	MH-NHP
rs237911	OTR	G/A	A/A	A/A	A/A	G/T	-
rs237917	OTR	T/C	C/C	C/C	C/C	C/C	MHU
rs10784339	VTR1A	G/C	G/C	G/G	G/G	G/G	HU
rs11174811	VTR1A	C/A	C/C	C/C	C/C	C/C	MHU
rs3803107	VTR1A	G/A	A/A	G/G	G/G	G/G	HS
rs1042615	VTR1A	A/G	G/G	G/G	A/G	G/G	MH-NHP
rs3021529	VTR1A	G/A	G/G	G/G	G/G	G/G	MHU
rs10877969	VTR1A	T/C	C/C	C/C	C/C	G/G	MHU
rs3759292	VTR1A	A/G	A/A	A/A	A/A	A/A	MHU
rs33985287	VTR1B	T/C	C/G	C/G	C/G	C/C	MHS
rs28676508	VTR1B	C/T	C/G	C/G	C/G	C/C	MHS
rs28632197	VTR1B	C/T	C/G	C/G	C/G	C/C	MHS

b

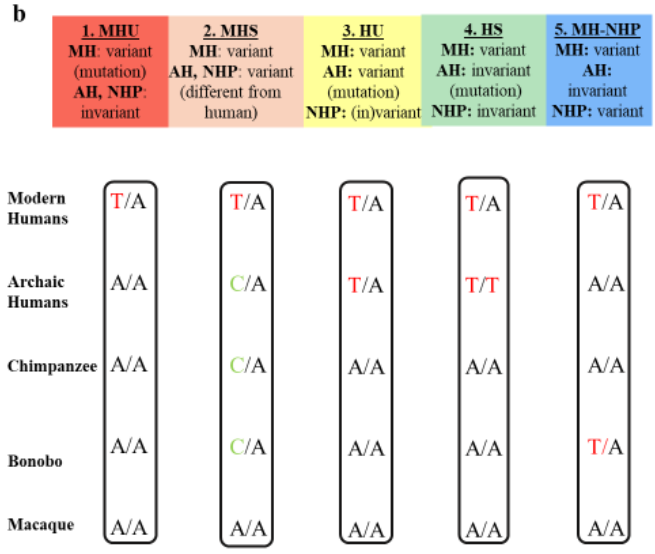

Figure 1: Variant clustering. a, For each variant site in *OTR*, *VTR1A* and *VTR1B*, we list the variation pattern found in MH (Modern Humans), AH (Archaic humans: Neanderthals and Denisovans), Chimpanzees, Bonobos and Macaques. In the column 'Type' we classify each allele on its assigned Type-category. Alleles that were key for each classification are in red font. **b,** A schematic representation of the four classification types. Shown in the columns are examples of each type. More details described for each type are in the main text results.

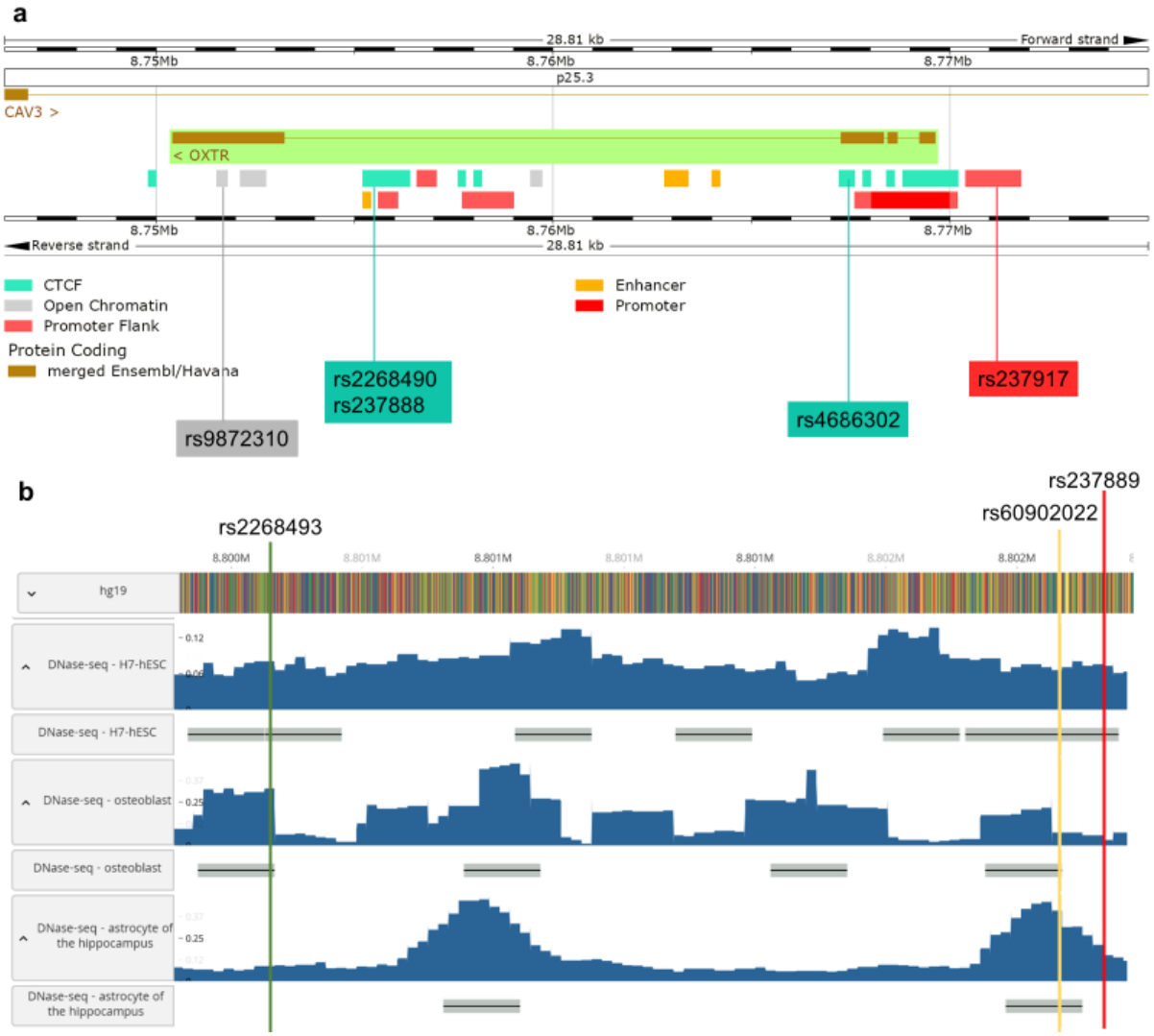

Figure 2: Ensembl regulation and RegulomeDB results. a, Ensembl regulation data resources for *OTR* (chromosome 3), showing the *CTCF* (transcription factor) binding domain (green), the open chromatin regions (grey), the promoter regions (red) and the enhancer regions (yellow). We highlighted SNP sites analyzed in this study. **b,** RegulomeDB browser (chromosome 3: 8,800-8,802 Mb, *OTR* intron) showing open chromatin regions (blue peaks) for 3 sites identified in this study. Shown are DNase sequencing data from undifferentiated embryonic stem cells (DNase-seq-H7-hESC), from osteoblasts, and from astrocytes from the hippocampus.

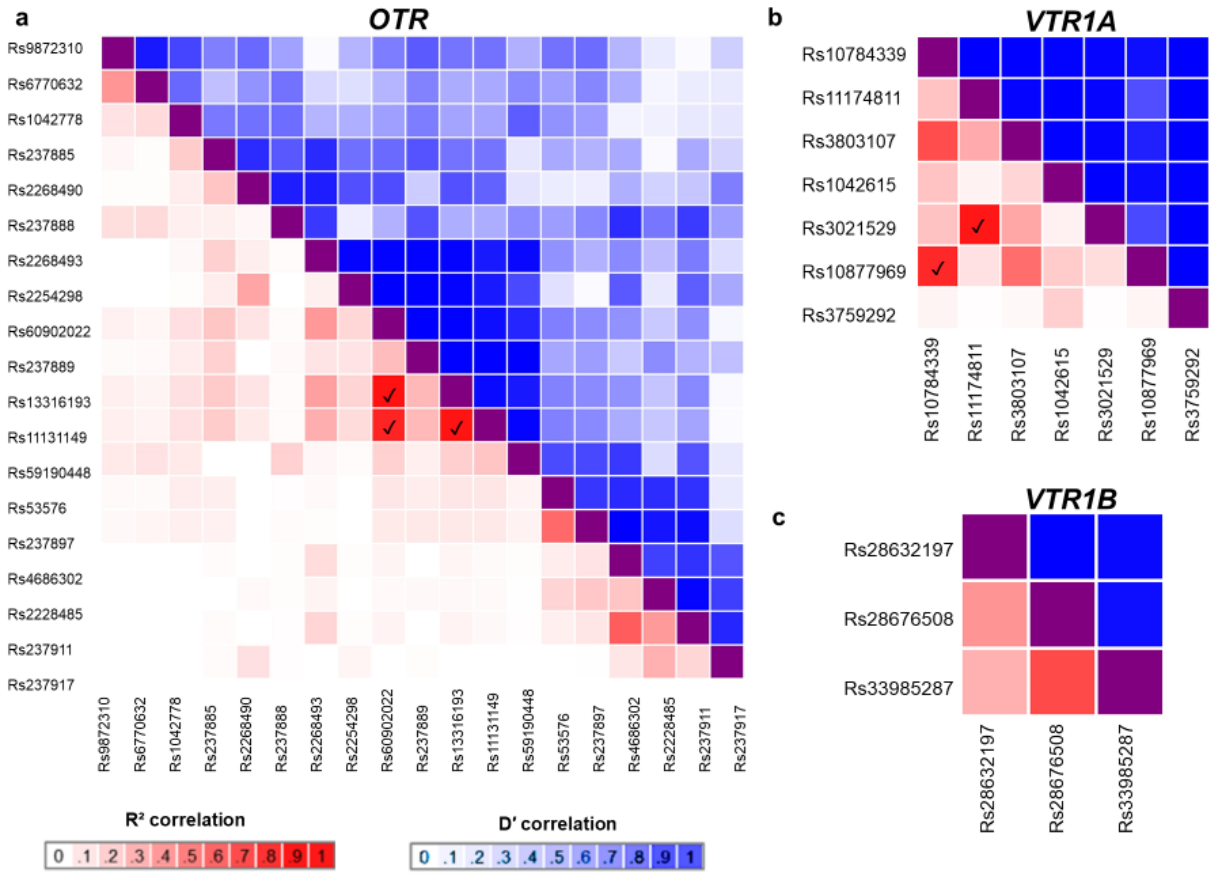

Figure 3: Linkage Disequilibrium analysis. a, *OTR* heatmap matrix of pairwise linkage disequilibrium statistics. Pairwise statistics are colored based on their R^2 and D´ significance scores, with the darkest shades of red being the most significant R^2 score, and white the least, and the darkest shades of blue being the highest D´ scores, and white the least. ✓, pairwise $R^2 > 0.8$, our threshold for a significant co-inheritance of the variant alleles. The same coloring and symbols were used in the rest of the heatmaps. **b,** *VTR1A* heatmap matrix of pairwise linkage disequilibrium statistics. **c,** *VTR1B* heatmap matrix of pairwise linkage disequilibrium statistics.

Variant ID	Consequence	Gene	Exon	Intron	Protein position	Amino acids	MH variants (Ancestral in bold)	Type	MAF	Highest MAF	CADD	GERP	Mutation Assessor	REVEL	Meta LR	SIFT	PolyPhen-2
rs9872310	3'-UTR	OTR	4/4	-	-	-	**A**/G	HS	G: 0.13	AFR: 0.34	G: 6.845	-0.13	-	-	-	-	-
rs6770632	3'-UTR	OTR	4/4	-	-	-	C/**A**	MHU	A: 0.23	AFR: 0.49	A: 3.168	-4.08	-	-	-	-	-
rs1042778	3'-UTR	OTR	4/4	-	-	-	G/A/C/**T**	MHU	T: 0.41	AMR: 0.50	A:0.162 C:0.133 T:0.119	-8.61	-	-	-	-	-
rs237885	intron	OTR	-	3/3	-	-	**T**/**G**	MHU	G: 0.49	SAS: 0.50	G:3.630	-0.87	-	-	-	-	-
rs2268490	intron	OTR	-	3/3	-	-	**C**/T	MHU	T: 0.26	EAS: 0.50	T:0.522	-2.78	-	-	-	-	-
rs237888	intron	OTR	-	3/3	-	-	T/**C**	HU	C: 0.13	AFR: 0.43	C:1.428	-2.46	-	-	-	-	-
rs2268493	intron	OTR	-	3/3	-	-	**T**/C	MHU	C: 0.20	AMI: 0.43	C:0.989	-3.54	-	-	-	-	-
rs2254298	intron	OTR	-	3/3	-	-	**G**/A	MHU	A: 0.21	EAS: 0.40	A:1.715	**0.57**	-	-	-	-	-
rs60902022	intron	OTR	-	3/3	-	-	**T**/C	HU	C: 0.39	AFR: 0.49	C:6.992	-2.93	-	-	-	-	-
rs237889	intron	OTR	-	3/3	-	-	T/**C**	MHU	T: 0.30	EAS: 0.48	C:2.691	-2.15	-	-	-	-	-
rs13316193	intron	OTR	-	3/3	-	-	**T**/C	HS	C: 0.40	AMI: 0.48	C:5.931	**1.4**	-	-	-	-	-
rs11131149	intron	OTR	-	3/3	-	-	**G**/A	-	A: 0.40	AMR: 0.49	A:1.550	-1.9	-	-	-	-	-
rs59190448	intron	OTR	-	3/3	-	-	G/**A**	HS	A: 0.14	AFR: 0.50	A:0.604	-2.55	-	-	-	-	-
rs53576	intron	OTR	-	3/3	-	-	A/**G**/T	MHU	A: 0.39	SAS: 0.49	G:3.423 T:2.950	-2.25	-	-	-	-	-
rs237897	intron	OTR	-	3/3	-	-	A/C/**G**	MH- NHP	A: 0.40	EUR: 0.50	C:1.870 G:2.090	-3.79	-	-	-	-	-
rs4686302	missense	OTR	3/4	-	218	A/T	**C**/T	HS	T: 0.14	AMR: 0.37	**T:19**	-	**0.089**	**0.013**	**0**	**0.4**	**0.01**
rs2228485	synonymous	OTR	3/4	-	57	N	G/**A**	MH- NHP	G: 0.28	AMR: 0.46	**A:20.4**	-0.47	-	-	-	-	-
rs237911	5'-UTR	OTR	3/4	-	-	-	G/**A**	-	G: 0.14	AMR: 0.38	A:5.562	-2.21	-	-	-	-	-
rs237917	upstream	OTR	-	-	-	-	T/**C**/G	MHU	T: 0.42	AMR: 0.50	C:4.859 G:4.557	-3.5	-	-	-	-	-
rs10784339	3'-UTR	VTR1A	2/2	-	-	-	**G**/A/C/T	HU	C: 0.31	AFR: 0.58	A:1.246 C:1.080 T:0.996	-1.02	-	-	-	-	-
rs11174811	3'-UTR	VTR1A	2/2	-	-	-	**C**/A	MHU	A: 0.10	SAS: 0.26	A:1.408	-2.23	-	-	-	-	-
rs3803107	3'-UTR	VTR1A	2/2	-	-	-	**G**/A	HS	A: 0.24	SAS: 0.46	A:2.640	-5.41	-	-	-	-	-
rs1042615	missense	VTR1A	1/2	-	136	F/L	A/C/**G**/T	MH- NHP	A: 0.35	EUR: 0.49	**C:28.6** **G:22.7** **T:28.5**	-2.04	**0.872**	**0.693**	**0.46**	**0**	**1**
rs3021529	intron	VTR1A	-	1/2	-	-	**G**/A	MHU	A: 0.10	SAS: 0.26	A:10.55	-4.07	-	-	-	-	-
rs10877969	intron	VTR1A	-	2/2	-	-	T/**C**	MHU	C: 0.30	AFR: 0.46	C:0.816	-2.9	-	-	-	-	-
rs3759292	intron	VTR1A	-	2/2	-	-	**A**/G	MHU	G: 0.10	EAS: 0.39	G:3.043	-1.93	-	-	-	-	-
rs33985287	3'-UTR	VTR1B	2/2	-	-	-	T/**C**	MHS	C: 0.29	AFR: 0.47	C:2.581	-	-	-	-	-	-
rs28676508	synonymous	VTR1B	2/2	-	373	S	**C**/G/T	MHS	T: 0.24	AFR: 0.50	G:0.443 T:0.551	-	-	-	-	-	-
rs28632197	missense	VTR1B	2/2	-	364	R/H	**C**/A/T	MHS	T: 0.12	SAS: 0.26	A:6.614, T:5.480	-	-	-	-	**0.27/ 0.07**	**0/0.001**

Table 1: List of identified variants in Modern humans and pathogenicity prediction scores. For each variant, we have listed where it is found in the genetic sequence (intronic, 3′-UTR) and if in the exon, whether it gives rise to a missense or a synonymous mutation ('Consequence'); on which specific 'Exon' or 'Intron' each site is located; on which 'Protein position' it is located, if exonic, and what 'Amino acids' it gives rise to; all MH (Modern human) alleles found on this site, with the ancestral being in bold; the 'Type' of each site based on our 5 categories (MHU: Modern Human Unique; MHS: Modern Human Specific; HU: Homo Unique; HS: Homo Specific; MH-NHP: Modern Human- Non-Human Primate); the Major Allele Frequency (MAF); the Highest Minor Allele Frequency (MAF) in a specific population from the 1000 Genomes or the gnomAD project (AFR: African, AMR: American, EAS: East Asian, SAS: South Asian, AMI: Amish);

CADD (Combined Annotation Dependent Depletion) and GERP (Genomic Evolutionary Rate Profiling), Mutation Assessor, REVEL (Rare Exome Variant Ensemble Learner), MetaLR (Meta Logistic Regression), SIFT (Sorting Tolerant From Intolerant) and PolyPhen-2 scores. Scores in red: deleterious; scores in yellow: neutral; scores in green: benign.

Variant ID	Gene	Ranking	ChIP	DNase	FAIRE	Ensemble	QTL	Proteins bound	Motifs altered (PWM)	Active in the brain
rs237889	OTR	1f	□	□	-		CAMK1	EGR2, ZFHX2	-	brain general
rs2228485	OTR	2b	□	□	□		-	CTCF, EZH2, POLR2A, RAD21, IKZF1	PAX6, ZSCAN4C	SN, Cd, CG, Hi, MFA 6, ANG, TL
rs2268493	OTR	3a	-	□	-		-	POLR2A	EWSR1, ONECUT3, POU3F1/2/3, REST, SPI1	LOW
rs60902022	OTR	3a	□	□	-		-	EGR2, ZFHX2	RAB11FIP5	brain general
rs237888	OTR	3a	□	□	-		-	*+CTCF, SOX13, FOXA3, GATAD2A	RFX7	CG
rs28632197	VTR1B	3a	□	□	-		-	POLR2A, ZBTB7B	TFAP2B	CG, Hi
rs2268490	OTR	4	□	□	-		-	*CTCF	-	CG
rs4686302	OTR	4	□	□	-		-	*CTCF	-	SN, Cd, CG, Hi, MFA 6, ANG, TL
rs237917	OTR	4	□	□	-	Promoter	-	-	-	SN, CG, Hi, MFA 6, ANG, TL
rs6770632	OTR	4	-	□	□		-	EZH2	-	CG
rs59190448	OTR	4	□	-	-		-	POLR2A	-	brain general
rs10877969	VTR1A	4	-	-	-		-	PKNOX1	-	SN, Cd, CG, Hi, MFA 6, ANG, TL
rs33985287	VTR1B	4	-	-	-		-	POLR2A	-	CG, Hi
rs1042615	VTR1A	4	-	□	-		-	EZH2, SUZ12, MAX, RNF2, ZBTB1	-	SN, CG, Hi, MFA 6, ANG, TL
rs9872310	OTR	5	-	□	-	Open chromatin	-	-	-	CG

Table 2: RegulomeDB high-ranked results and Ensembl regulation results. For each site, we list the gene they are found in, ranking scores from RegulomeDB, if they are found at open chromatin regions based on Chip-seq, DNase-seq or FAIRE-seq (□) data from Ensembl regulation database. This includes sites found on promoter or open chromatin regions, expression quantitative trait loci (eQTL) regions, proteins bound (* for those mentioned only in Ensembl, *+ for those supported from both RegulomeDB and Ensembl data. All the rest are come from RegulomeDB), PWM (positional weight matrix), whether they are active in the brain (e.g. in chromatin with evidence for strong transcription in the brain) and in which brain regions (SN: Substantia nigra; Cd: Caudate nucleus; CG: Cingular gyrus; Hi: Hippocampus; MFA 46: Middle Frontal Area 46; ANG: Angular Gyrus; TL: temporal lobe).

	FDIST			BayeScan				FLK		
rsID	H_E	F_{ST}	P	**log10(BF)**	**q-val**	α_i	F_{ST}	**Ht**	**F.LK**	P
rs6770632	0.3567	0.1545	0.9551	-0.3203	0.2942	0.2607	0.1131	0.423	3.988	0.136
rs1042778	0.4776	0.1635	0.9851	-1.2907	0.8133	0.0161	0.0891	0.467	0.142	0.932
rs237885	0.5013	0.0555	0.2712	0.1385	0.1806	-0.6004	0.0581	0.499	0	1
rs237888	0.196	0.1925	0.798	-0.6614	0.4348	0.1349	0.1009	0.267	8.993	0.011
rs59190448	0.1825	0.3103	0.0004	3.3977	0.0004	1.447	0.2548	0.262	13.521	0.001
rs10877969	0.366	0.1935	0.8974	-1.3998	0.8156	0.0046	0.0738	0.377	2.584	0.275
rs11174811	0.1467	0.0047	0	1.2742	0.0336	-1.3077	0.0255	0.136	0.622	0.733
rs3021529	0.1567	0.0057	0.0001	1.2035	0.0399	-1.2604	0.0266	0.147	0.594	0.743
rs3759292	0.1552	0.2003	0.1136	-0.7408	0.3899	0.1121	0.0832	0.176	14.129	0.001
rs237911	0.2718	0.0315	0.0485	-1.1798	0.7611	-0.0238	0.0864	0.266	2.701	0.259
rs2228485	0.4217	0.0238	0.0149	-1.1187	0.7285	-0.0313	0.086	0.437	3.486	0.175

Extended Table 1: Selection analysis based on Schaschl et al. 2015[1] results. Listed are the results from the FDIST, Bayescan and Lewontin and Krackauer extended tests of our identified SNPs found in the Schaschl et al. 2015[1] Additional File. The results with significance based on our filtering (P or q value□<□0.05) are highlighted in green (: positive selection) or yellow (: balancing selection). In FDIST: H_E, expected heterozygosity; F_{ST}, locus-specific genetic divergence among human populations. In BayeScan (User Manual; http://cmpg.unibe.ch/software/BayeScan/files/BayeScan2.1_manual.pdf): log10(BF)□> 1 indicate 'strong' evidence for selection and >1.5 indicate 'very strong' evidence for positive selection; q-val, the BayeScan analogue of a P value; $\alpha_i < 0$, balancing or purifying selection; $\alpha_i > 0$; diversifying selection ($\alpha_i > 0$); F_{ST}, locus-specific genetic divergence among human populations. In FLK: Ht, estimated heterozygosity,

Lewontion and Krackauer extended test statistics (F.LK) and P values for F.LK.F_{ST} values differ due to different calculation algorithms.

Acknowledgements
We thank Thomas O'Rourke for comments on the manuscript. We also thank Evan E. Eichler for guidance on primate variation data.

Funding statement
CB acknowledges the financial support from the Spanish Ministry of Economy and Competitiveness/FEDER funds (grant FFI2016-78034-C2-1-P), the Spanish Ministry of Science and Innovation (PID2019-107042GB-I00), a Marie Curie International Reintegration Grant from the European Union (PIRG-GA-2009-256413), research funds from the Fundacio´ Bosch i Gimpera, from the Generalitat de Catalunya (2017-SGR-341), and the MEXT/JSPS Grant-in-Aid for Scientific Research on Innovative Areas 4903 (Evolinguistics: JP17H06379). CTh acknowledges support from the Generalitat de Catalunya in the form of a doctoral (FI) and the Rockefeller University. AA acknowledges financial support from the Spanish Ministry of Economy and Competitiveness and the European Social Fund (BES-2017-080366). EDJ is funded by the Howard Hughes Medical Institute and the Rockefeller University.

Author Contributions Statement
CTh conceptualized and designed the study, ran variation, pathogenicity prediction, regulation, human population and macaque data, and association analyses. AA ran variation, pathogenicity prediction, human, chimpanzee and bonobo variation analyses and reviewed the literature. EDJ and CB supervised the study.

Data availability statement
All data generated or analyzed during this study are included in the Suppl. Info and Tables.

Competing interests' statement
There is NO Competing financial or non-financial interest.

References

1. Schaschl, H. & others. Signatures of positive selection in the cis-regulatory sequences of the human oxytocin receptor (OXTR) and arginine vasopressin receptor 1a (AVPR1A) genes. *BMC Evol. Biol.* **15**, 85 (2015).
2. Dunbar, R. I. M. & Shultz, S. Evolution in the Social Brain. *Science.* **317**, 1344–1347 (2007).
3. Tomasello, M. The ultra-social animal. *Eur. J. Soc. Psychol.* **44**, 187–194 (2014).
4. Rogers, A. R., Harris, N. S. & Achenbach, A. A. Neanderthal-Denisovan ancestors interbred with a distantly related hominin. *Sci. Adv.* **6**, eaay5483 (2020).
5. Prüfer, K. & others. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
6. Meyer, M. & others. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science.* **338**, 222–226 (2012).
7. Castellano, S. *et al.* Patterns of coding variation in the complete exomes of three Neandertals. *Proc. Natl. Acad. Sci.* **111**, 6666–6671 (2014).
8. Rogers, A. R., Bohlender, R. J. & Huff, C. D. Early history of Neanderthals and Denisovans. *Proc. Natl. Acad. Sci.* **114**, 9859–9863 (2017).
9. Donaldson, Z. R. & Young, L. J. Oxytocin, Vasopressin, and the Neurogenetics of Sociality. *Science.* **322**, 900–904 (2008).
10. Meyer-Lindenberg, A., Domes, G., Kirsch, P. & Heinrichs, M. Oxytocin and vasopressin in the human brain: Social neuropeptides for translational medicine. *Nat. Rev. Neurosci.* **12**, 524–538 (2011).
11. Israel, S. & others. Molecular genetic studies of the arginine vasopressin 1a receptor (AVPR1a) and the oxytocin receptor (OXTR) in human behaviour: from autism to altruism with some notes in between. *Elsevier* **170**, 435–449 (2008).
12. Theofanopoulou, C., Gedman, G., Cahill, J. A., Boeckx, C. & Jarvis, E. D. Universal

nomenclature for oxytocin-vasotocin ligand and receptor families. *Nature* (accepted).
13. Hare, B. Survival of the Friendliest: Homo sapiens Evolved via Selection for Prosociality. *Annu. Rev. Psychol.* **68**, 155–186 (2017).
14. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
15. Prüfer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527–531 (2012).
16. Kuhlwilm, M. & Boeckx, C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci. Rep.* **9**, 8463 (2019).
17. Prüfer, K. & others. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science.* (2017) doi:10.1126/science.aao1887.
18. Semal, P. *et al.* New data on the late Neandertals: Direct dating of the Belgian Spy fossils. *Am. J. Phys. Anthropol.* **138**, 421–428 (2009).
19. Rougier, H. *et al.* Neandertal cannibalism and Neandertal bones used as tools in Northern Europe. *Sci. Rep.* **6**, 29005 (2016).
20. Hajdinjak, M. *et al.* Reconstructing the genetic history of late Neanderthals. *Nature* **555**, 652–656 (2018).
21. Peyrégne, S. *et al.* Nuclear DNA from two early Neandertals reveals 80,000 years of genetic continuity in Europe. *Sci. Adv.* **5**, eaaw5873 (2019).
22. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci.* **117**, 15132–15136 (2020).
23. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *bioRxiv* 2020.03.12.988956 (2020) doi:10.1101/2020.03.12.988956.
24. Hammock, E. A. D. Microsatellite Instability Generates Diversity in Brain and Sociobehavioral Traits. *Science.* **308**, 1630–1634 (2005).
25. Donaldson, Z. R. *et al.* Evolution of a behavior-linked microsatellite-containing element in the 5' flanking region of the primate AVPR1A gene. *BMC Evol. Biol.* **8**, 180 (2008).
26. Lerer, E. *et al.* Association between the oxytocin receptor (OXTR) gene and autism: Relationship to Vineland Adaptive Behavior Scales and cognition. *Mol. Psychiatry* **13**, 980–988 (2008).
27. Campbell, D. B. & others. Association of oxytocin receptor (OXTR) gene variants with multiple phenotype domains of autism spectrum disorder. *J. Neurodev. Disord.* **3**, 101–112 (2011).
28. Ribeiro, L. de O. P. *et al.* Evidence for Association Between OXTR Gene and ASD Clinical Phenotypes. *J. Mol. Neurosci.* 1–9 (2018).
29. Israel, S. & others. The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the Dictator Game and the social value orientations task. *PLoS One* **4**, (2009).
30. Wu, S. & others. Positive association of the oxytocin receptor gene (OXTR) with autism in the Chinese Han population. *Biol. Psychiatry* **58**, 74–77 (2005).
31. Feldman, R. & others. Sensitive Parenting Is Associated with Plasma Oxytocin and Polymorphisms in the OXTR and CD38 Genes. *Biol. Psychiatry* **72**, 175–181 (2012).
32. Dobewall, H. *et al.* Oxytocin receptor gene (OXTR) variant rs1042778 moderates the influence of family environment on changes in perceived social support over time. *J. Affect. Disord.* **235**, 480–488 (2018).
33. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
34. Cooper, G. M. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
35. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118–e118 (2011).
36. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
37. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137

(2015).
38. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
39. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
40. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
41. Zerbino, D. R. *et al.* Ensembl regulation resources. *Database* **2016**, bav119 (2016).
42. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants: Fig. 1. *Bioinformatics* **31**, 3555–3557 (2015).
43. Auton, A. & others. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
44. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
45. Kovalaskas, S., Rilling, J. K. & Lindo, J. Comparative analyses of the Pan lineage reveal selection on gene pathways associated with diet and sociality in bonobos. *Genes, Brain Behav.* (2020) doi:10.1111/gbb.12715.
46. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471 (2013).
47. Luzzatto, L. Sickle cell anaemia and malaria. *Mediterr. J. Hematol. Infect. Dis.* **4**, e2012065 (2012).
48. Baribeau, D. A. *et al.* Oxytocin receptor polymorphisms are differentially associated with social abilities across neurodevelopmental disorders. *Sci. Rep.* **7**, 11618 (2017).
49. Apps, M. A. J., Rushworth, M. F. S. & Chang, S. W. C. The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. *Neuron* **90**, 692–707 (2016).
50. Gu, X. *et al.* Interplay of miR-137 and EZH2 contributes to the genome-wide redistribution of H3K27me3 underlying the Pb-induced memory impairment. *Cell Death Dis.* **10**, 671 (2019).
51. Li, J. *et al.* Chromatin remodeling gene EZH2 involved in the genetic etiology of autism in Chinese Han population. *Neurosci. Lett.* **610**, 182–186 (2016).
52. Choufani, S. *et al.* DNA Methylation Signature for EZH2 Functionally Classifies Sequence Variants in Three PRC2 Complex Genes. *Am. J. Hum. Genet.* **106**, 596–610 (2020).
53. Theofanopoulou, C. *et al.* Self-domestication in homo sapiens: Insights from comparative genomics. *PLoS One* **12**, 10 (2017).
54. Theofanopoulou, C. Implications of oxytocin in speech. PhD Thesis. University of Barcelona, 2019.
55. Zanella, M. *et al.* Dosage analysis of the 7q11.23 Williams region identifies BAZ1B as a major human gene patterning the modern human face and underlying self-domestication. *Sci. Adv.* **5**, eaaw7908 (2019).
56. Wrangham, R. *The goodness paradox: the strange relationship between virtue and violence in human evolution.* (Vintage, 2019).
57. Fam, B. S. O. *et al.* Oxytocin and arginine vasopressin systems in the domestication process. *Genet. Mol. Biol.* **41**, 235–242 (2018).
58. Bence, M. *et al.* Lessons from the canine Oxtr gene: populations, variants and functional aspects. *Genes, Brain Behav.* **16**, 427–438 (2017).
59. Ruan, C. & Zhang, Z. Laboratory domestication changed the expression patterns of oxytocin and vasopressin in brains of rats and mice. *Anat. Sci. Int.* **91**, 358–370 (2016).
60. Dunbar, R. I. M. *Human evolution*. (Oxford University Press, 2016).
61. Raghanti, M. A. & others. A neurochemical hypothesis for the origin of hominids. *Proc. Natl. Acad. Sci.* 201719666 (2018) doi:10.1073/pnas.1719666115.
62. Hunt, B. G. *et al.* Relaxed selection is a precursor to the evolution of phenotypic plasticity. *Proc. Natl. Acad. Sci.* **108**, 15936–15941 (2011).
63. Tan, J., Ariely, D. & Hare, B. Bonobos respond prosocially toward members of other groups. *Sci. Rep.* **7**, (2017).
64. Hare, B. & Wrangham, R. W. Equal, Similar, but Different: Convergent Bonobos and

Conserved Chimpanzees. in *Chimpanzees and Human Evolution* (eds. Muller, M. M., Wrangham, R. & Pilbeam, D. R.) 142–173 (Harvard University Press, 2017).
65. Pearce, E., Wlodarski, R., Machin, A. & Dunbar, R. I. M. Variation in the beta-endorphin, oxytocin, and dopamine receptor genes is associated with different dimensions of human sociality. *Proc. Natl. Acad. Sci.* **114**, 5300–5305 (2017).
66. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
67. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
68. Brown, N. P., Leroy, C. & Sander, C. MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics* **14**, 380–381 (1998).
69. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
70. Wright, E. DECIPHER. (2017) doi:10.18129/b9.bioc.decipher.
71. Esteller-Cucala, P. *et al.* Genomic analysis of the natural history of attention-deficit/hyperactivity disorder using Neanderthal and ancient Homo sapiens samples. *Sci. Rep.* **10**, 8622 (2020).
72. Staes, N. & others. Oxytocin and Vasopressin Receptor Gene Variation as a Proximate Base for Inter- and Intraspecific Behavioral Differences in Bonobos and Chimpanzees. *PLoS ONE* **9**, e113364 (2014).
73. Staes, N. & others. Bonobo personality traits are heritable and associated with vasopressin receptor gene 1a variation. *Sci. Rep.* **6**, 38193 (2016).
74. Staes, N. & others. Chimpanzee sociability is associated with vasopressin (Avpr1a) but not oxytocin receptor gene (OXTR) variation. *Horm. Behav.* **75**, 84–90 (2015).
75. Bimber, B. N., Yan, M. Y., Peterson, S. M. & Ferguson, B. mGAP: the macaque genotype and phenotype resource, a framework for accessing and interpreting macaque variant data, and identifying new models of human disease. *BMC Genomics* **20**, 176 (2019).
76. Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
77. NCBI Resource Coordinators. Database Resources of the NCBI. *Nucleic Acids Res.* **45**, D12--D17 (2016).