Diverse Functions Associate With Non-Coding Trans-species Polymorphisms in Humans

Keila Velazquez-Arcelay
Vanderbilt University
Mary Lauren Benton
Baylor University
John A. Capra (tony@capralab.org)
University of California, San Francisco

Research Article

Keywords: trans-species polymorphisms, balancing selection, long-term balancing selection, non-coding variants, phenome-wide association study

Posted Date: May 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-559297/v1

Diverse functions associate with non-coding trans-species polymorphisms in humans

Keila Velazquez-Arcelay[1], Mary Lauren Benton[2], and John A. Capra[1,3,4*]

[1] Department of Biological Sciences, Vanderbilt University
[2] Department of Computer Science, Baylor University
[3] Departments of Biomedical Informatics and Computer Science, Genetics Institute, and Center for Structural Biology, Vanderbilt University
[4] Bakar Computational Health Sciences Institute and Department of Epidemiology and Biostatistics, University of California, San Francisco
[*] Corresponding author: tony@capralab.org

Abstract

Background: Long-term balancing selection (LTBS) can maintain allelic variation at a locus over millions of years and through speciation events. Variants shared between species, hereafter "trans-species polymorphisms" (TSPs), often result from LTBS due to host-pathogen interactions. For instance, the major histocompatibility complex (MHC) locus contains TSPs present across primates. Several hundred candidate TSPs have been identified in humans and chimpanzees; however, because many are in non-coding regions of the genome, the functions and adaptive roles for most TSPs remain unknown.

Results: We integrated diverse genomic annotations, with a focus on non-coding regions, to explore the functions of 125 previously identified regions containing multiple TSPs in humans and chimpanzees. We analyzed genome-wide functional assays, expression quantitative trait loci (eQTL), genome-wide association studies (GWAS), and phenome-wide association studies (PheWAS). We identify functional annotations for 119 TSP regions, including 71 with evidence of gene regulatory function from GTEx or genome-wide functional genomics data and 21 with evidence of trait association from GWAS and PheWAS. TSPs in humans associate with many immune system phenotypes, including response to pathogens, but we also find associations with a range of other phenotypes, including body mass, alcohol intake, urate levels, chronotype, and risk-taking behavior.

Conclusions: The diversity of traits associated with non-coding human TSPs further support previous hypotheses that functions beyond the immune system are subject to LTBS. Furthermore, several of these trait associations provide support and candidate genetic loci for previous hypothesis about behavioral diversity in great ape populations, such as the importance of variation in sleep cycles and risk sensitivity.

Keywords: trans-species polymorphisms; balancing selection; long-term balancing selection; non-coding variants; phenome-wide association study

Significance statement

Most genetic variants present in human populations are young (<100,000 years old); however, a few hundred are millions of years old with origins before the divergence of humans and chimpanzees. Many of these trans-species polymorphisms (TSPs) were likely maintained by balancing selection—evolutionary pressure to maintain genetic diversity at a locus. In spite of their age, the functions driving this selection, especially for non-coding TSPs, are largely unknown. We integrate genome-wide annotation strategies to demonstrate TSP associations with immune system function, behavior (addition, cognition, risky behavior), uric acid metabolism, and many other phenotypes. These results substantially expand our understanding of functions for TSPs and support a role for balancing selection beyond the immune system.

Background

The interaction between populations and environments is dynamic. Over time, allele frequencies in a population shift due to drift and adaptive responses to specific environmental pressures. Most genetic variants are short-lived compared to the timescale of species. But on rare occasions variants persistently segregate at intermediate frequencies for millions of years, sometimes predating the most recent common ancestor (MRCA) between two sister species (Bitarello et al., 2018; Cheng & DeGiorgio, 2019; DeGiorgio et al., 2014; Leffler et al., 2013; Siewert & Voight, 2017; Teixeira et al., 2015). These trans-species polymorphisms (TSPs) are likely a sign of genomic regions under long-term balancing selection (LTBS). Over time, instances of LTBS leave signatures in the genome that differentiate them from those under other forms of selection (Bitarello et al., 2018; Key et al., 2014; Leffler et al., 2013; Siewert & Voight, 2017).

Several instances of likely LTBS have been observed in humans and other primates, mostly within the major histocompatibility complex (MHC) or the ABO blood group locus. For example, the MHC, or human leukocyte antigen (HLA) system in humans, is a family of varied proteins expressed on the cell surface with essential functions in adaptive immune response and regulation. Balancing selection on different components of the HLA region dates to the common ancestor between chimpanzees and humans (Lawlor et al., 1988; Mayer et al., 1988) (Azevedo et al., 2015). Similarly, the ABO gene has three alleles, and its variants lead to different blood cell antigens, or lack of thereof, on the surface of the cell. Variation in this group could have a benefit in the immune response to pathogens, and balanced polymorphisms at this locus date back to the common ancestor of gorillas, orangutans, and humans (Ségurel et al., 2012). Several other immune-related genes show LTBS between humans and other primates, e.g.: *TRIM5*, a RING finger protein 88 (Battivelli et al., 2011; Cagliani et al., 2010; Ganser-Pornillos & Pornillos, 2019), and *ZC3HAV1*, a zinc finger CCCH-type antiviral protein 1 (Cagliani et al., 2012; De Filippo et al., 2016; Mao et al., 2013; Todorova et al., 2015). These genes have important roles in host/pathogen response through inhibition of virus replication.

The high allelic variation maintained by balancing selection at a locus can also enable adaptive selection in new environments. For example, some variants found under balancing

selection in African and ancestral human populations have experienced directional selection in non-African populations (European and Asian), with one allele becoming predominant in the population (De Filippo et al., 2016). This suggests the adaptive potential of the variation maintained under balancing selection; however, in some cases the adaptive variants themselves may have hitchhiked with those under LTBS.

Recent studies have developed methods to identify instances of balancing selection in genome-wide data (Bitarello et al., 2018; Cheng & DeGiorgio, 2019; DeGiorgio et al., 2014; Siewert & Voight, 2017, 2020). Some have focused on detecting LTBS using trans-species data, while others have considered balancing selection over shorter timescales based on single-species data. For example, De Giorgio (2014) developed likelihood-ratio tests (T_1 and T_2) based on computing probabilities of polymorphism and substitution under LTBS based on inter-species coalescent modeling. With this method they identified balancing selection on HLA regions, but also in a gene that had no previous associations with balancing selection, *FANK1*, which is involved in the suppression of apoptosis during/after the process of meiosis. They also found enrichment in other functions: cell adhesion, membrane protein activity, and components of membranes. A more recent study (Cheng & DeGiorgio, 2019) expanded the T_2 method to seek trans-species balancing selection without direct consideration of trans-species polymorphism and identified a handful of additional LTBS candidates. Bitarello et al. (2018) developed Non-central Deviation (NCD) statistics that quantify the deviation from expectations under balancing selection of the local site frequency spectrum (SFS) in windows in African and European 1000 Genomes populations. They identified thousands of candidates for balancing selection in humans. They also showed varying directional selection in different populations, providing evidence for the adaptive potential of regions under balancing selection. Siewert & Voight (2017) developed ß, a summary statistic for detecting genomic windows with clusters of intermediate frequency alleles suggestive of balancing selection. They also recently updated the ß statistic to consider both polymorphism and substitution data (Siewert & Voight, 2020). Among the highest scoring windows in these two analyses, they highlighted three genes (*CADM2*, *WFS1*, and *ACSBG2*) with functions outside the immune system.

Trans-species polymorphisms, especially when more than one falls on a haplotype, suggest the action of LTBS. Leffler et al. (2013) compared polymorphisms across the genome in Yoruba individuals from the 1000 Genomes Project to those found in Western chimpanzees sequenced by the PanMap Project. They identified more than 100 non-coding haplotypes with multiple TSPs within 4 kb and in high LD suggesting the presence of LTBS. Sequencing errors and high mutation rates can produce TSPs (Cheng & DeGiorgio, 2019; Gao et al., 2015) but it is unlikely to observe haplotypes with two TSPs in close proximity by chance.

Despite the importance and prevalence of balancing selection, most the non-coding haplotypes bearing potential signatures of LTBS, i.e., multiple TSPs, have not been functionally characterized. Here, we focus on the non-coding TSPs identified by Leffler et al. (2013). Determining the functional roles of these polymorphisms in human adaptation and health will deepen our understanding of the dynamics of balancing and positive selection and their roles in

adaptation to new environments. We identify potential functions of the TSPs in humans by applying several genome-wide functional annotations and association tests. Our results identify diverse functions, including effects unrelated to the immune system, that may underlie LTBS on the human and chimpanzee lineages.

Results

Human-chimpanzee TSPs

We consider 125 human genomic regions containing multiple TSPs segregating in both humans and chimpanzees in close proximity and high LD from Leffler et al. (2013). This set is composed of 263 variants with strong evidence of identity-by-descent; i.e., the divergence between the ancestral and derived alleles is deeper than the human-chimp speciation event. These TSPs were identified based on the observation from coalescent theory (Ségurel et al., 2012) that pairs of TSPs within 4 kb in the human genome are extremely unlikely to result from neutral processes, and thus are strong candidates for LTBS (Figure 1). Hereafter, we refer to these as "TSP regions". While these criteria do not guarantee that all the TSP regions are the result of LTBS (Gao et al., 2015), this set is likely strongly enriched for LTBS. We also quantify the evidence for balancing selection on these loci provided by BetaScan2 and NCD, two additional recent genome-wide scans (Bitarello et al., 2018; Siewert & Voight, 2020). In total, 80% (100/125) of the TSP regions had additional evidence of balancing selection from at least one of BetaScan2 or NCD (Methods; Supplementary Table 7.

In the following, we analyze two sets of variants for the 125 TSP regions. First, we focus on the 263 TSPs themselves. Second, to capture functions tagged by variants in high linkage disequilibrium (LD) with TSPs, we also considered potential tag SNPs in high LD ($R^2 \geq 0.8$) with TSPs in African, European, or East Asian populations from the 1000 Genomes Project. This LD-expanded set includes 10,259 variants across the 125 TSP regions (Supplementary Figure 1). By expanding to include LD, we capture additional associations, but may also introduce false positives; thus, we report results on both sets throughout.

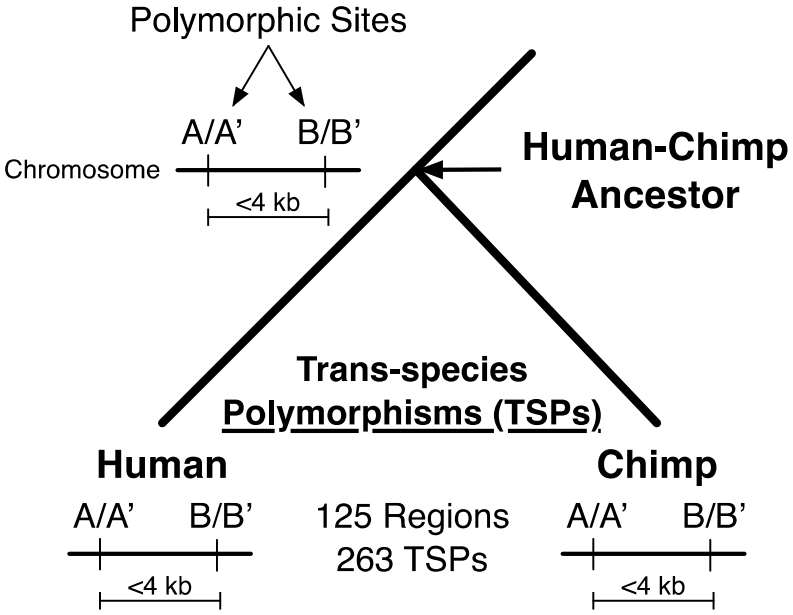

Figure 1. Trans-species polymorphisms (TSPs) likely resulting from long-term balancing selection (LTBS). Schematic showing the criteria used by Leffler et al. (2013) to identify TSPs likely maintained by LTBS. Each line represents a chromosome with polymorphisms segregating in a population. A/A' are two alleles segregating in both humans and chimpanzees at one site (i.e., a TSP), and B/B' are alleles segregating in both species at a nearby site. TSPs are very unlikely to appear nearby (within 4 kb) without the action of balancing selection. We consider 125 TSP regions containing 263 TSPs. Within these regions, multiple functional scenarios are possible. For example, one TSP may be under LTBS while the other is neutral, but maintained due to tight linkage. Alternatively, the TSPs may have epistatic functions and both be under selection. In addition to the 263 TSPs, we also considered functional associations with 9,996 variants in high LD ($r^2 > 0.8$) with a TSP at least one population from the 1000 Genomes Project (Supplementary Figure 1).

Trans-species polymorphisms overlap diverse functional annotations

We intersected the TSPs with diverse lines of functional evidence from large-scale genomic studies, including genome-wide functional genomics assays, eQTL, GWAS, and PheWAS. We found at least one functional annotation for 95% of the TSP regions (119 out of 125) covering 130 TSPs and 4,807 LD SNPs (Figure 2). Here, we provide an overview of the overlap with these annotations. In future sections, we provide details about each of these annotations. Variants in 91% (114 out of 125) of regions overlap annotated gene regulatory regions. This includes 58 TSPs in 40 regions and 1334 LD variants in 112 regions. We also found 86 TSPs across 51 regions with evidence of being expression quantitative trait loci (eQTL) in 29 tissues. Including the variants in LD with TSPs, 57.6% of the regions (72 out of 125) contain eQTLs ($p < E\text{-}5$) in

49 tissues. We found genome-wide significant associations with phenotypes in available genome- or phenome-wide association studies for 19% of the regions (24 out of 125; 13 GWAS and 16 geneAtlas PheWAS). Finally, 11.2% (14 out of 125) of these regions contain SNPs in protein-coding regions; one TSP (rs12918619) produces a non-synonymous protein sequence change, and 16 other non-synonymous variants in 7 regions are in high LD with the TSPs. (These were not identified as coding in the previous study due to changes in protein annotations.)

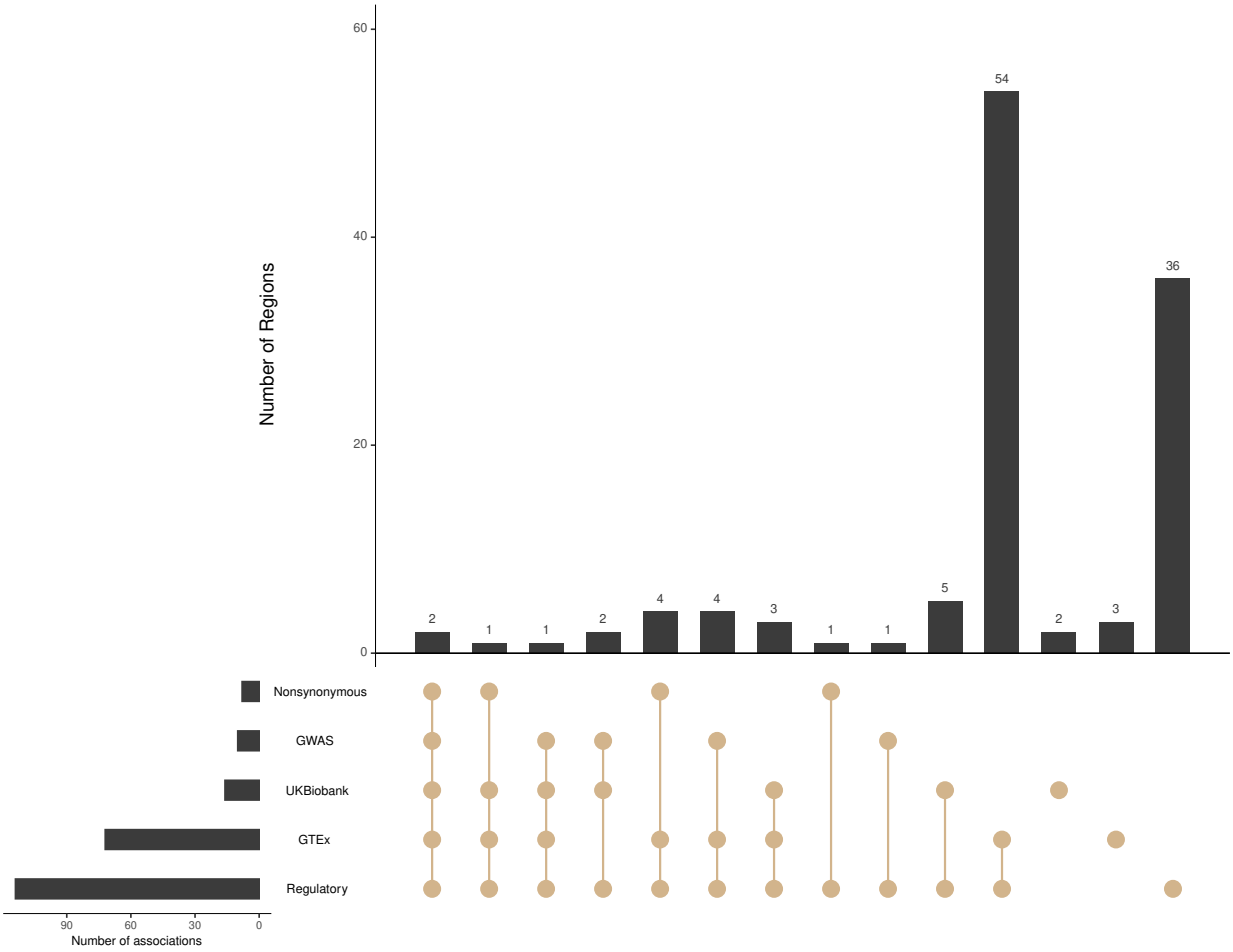

Figure 2. Functional annotations available for the expanded TSP regions. Summary of the annotations of each type available for TSP regions, including tagging variants in high LD with TSPs. A total of 119 out of 125 TSP regions contain at least one line of functional evidence. Multiple lines (two or more) are available for 78 regions.

Evidence of gene regulatory function for TSPs

We hypothesized that many of the non-coding TSPs in our set perform gene regulatory functions. To evaluate this possibility, we intersected the TSPs and variants in high LD with maps of functional regulatory regions from the Ensembl regulatory build (Zerbino et al., 2015). We found 58 TSPs with regulatory annotations in 40 TSP regions, and additionally 1334 LD

variants in 114 regions. These include variants in CTCF binding sites, open chromatin regions, promoter flanking regions, enhancers, promoters, and known TF binding sites (Supplementary Table 1).

Overlap of a variant with a regulatory annotation does not necessarily imply a regulatory function. To consider additional evidence of regulatory function, we examined eQTL in GTEx from 50 tissues for overlap with TSPs. At least one eQTL was found for 51 of the TSP regions (40%). Among these 51 regions, 29 TSPs are themselves eQTL; for the remainder, variants in high LD with TSPs were eQTL. The eQTL were found across diverse tissues (Figure 3A), and there was no enrichment for specific gene ontology (GO) terms among the set of genes influenced by TSP eQTL. This suggests that the targets of balancing selection have functions in gene regulation across diverse tissues beyond the immune system.

Next, we tested if the TSP regions are enriched for overlap with any specific types of regulatory regions. We compared the observed overlap between TSP regions and each type of regulatory annotation to the distribution of overlaps expected if TSP regions were randomly distributed across the genome. We shuffled the TSP regions 1,000 times maintaining their length and chromosome distributions and avoiding genome assembly gaps and ENCODE blacklist regions and counted the number of overlaps observed with regulatory elements for each random permutation. The TSPs overlap slightly fewer base pairs annotated with regulatory functions than expected by chance, with significant ($P < 0.05$) depletion for promoter and CTCF sites (Figure 3B). Since variants in these regions are likely to influence gene regulation in many tissues (e.g., compared to enhancers which are often context-specific), this suggests that individual TSPs may be less pleiotropic than expected by chance.

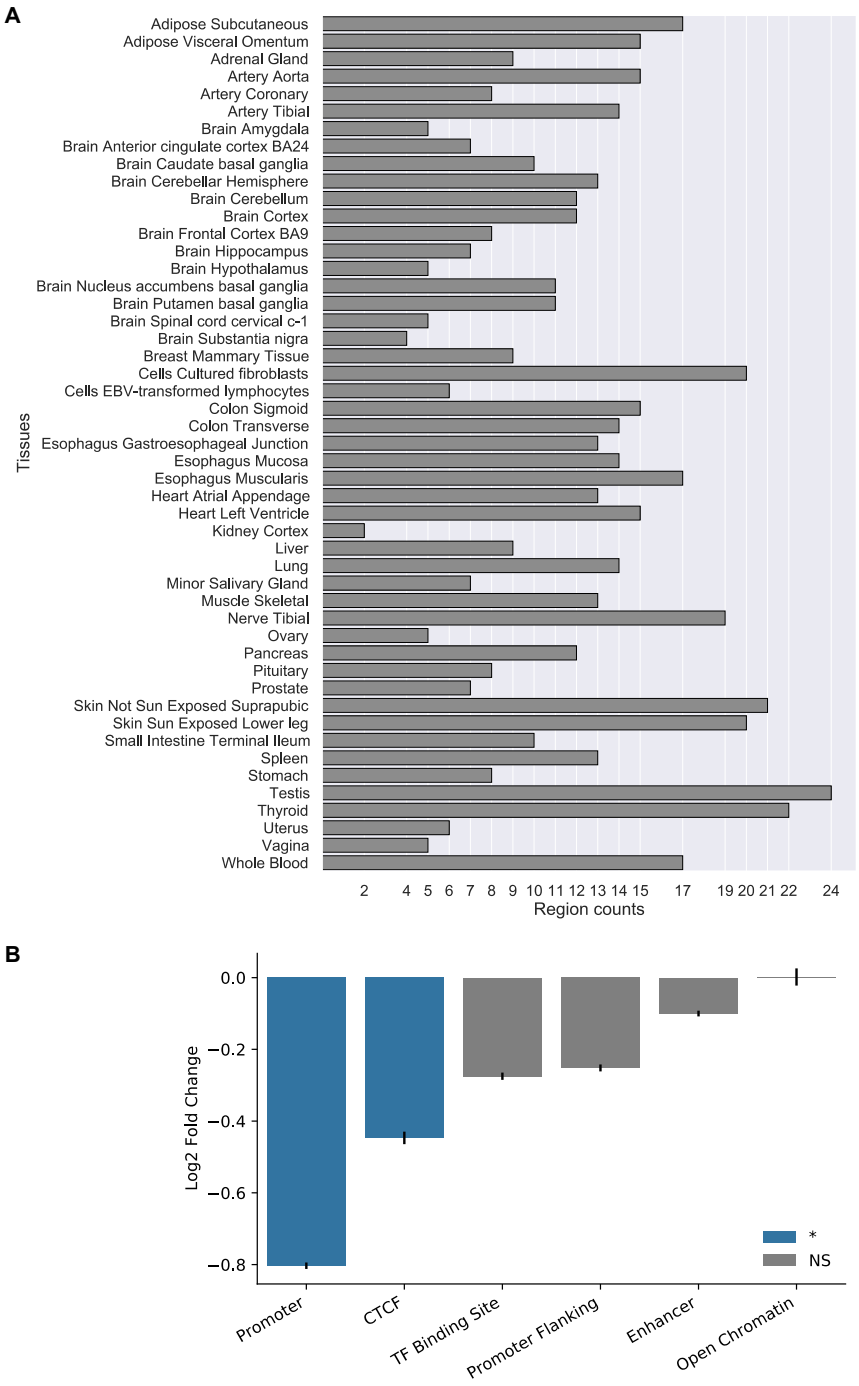

Figure 3. TSPs are eQTLs in diverse tissues and are depleted for overlap with promoters and CTCF sites. (A) The number of TSP regions that contain an eQTL for each GTEx tissue. Variation in TSP regions associates with gene expression in diverse tissues. The associated genes also have Gene Ontology (GO) annotations from diverse functional categories (Supplementary Figure 2). (B) TSP regions are significantly depleted for overlap with promoters and CTCF sites compared to length- and chromosome-matched non-coding regions from the genomic background. The error bars represent 95% confidence intervals.

Genome-wide association studies link TSPs to traits

Genome-wide association studies have identified thousands of associations between genetic variants and human traits. We intersected the TSP regions with associations reported in the GWAS Catalog (downloaded 2020/11), which is composed of 227,262 associations. Since TSPs themselves were not always directly tested in GWAS studies, we also include genome-wide significant ($p < 1E-8$) associations with the tag variants in high LD with TSPs. We found significant associations for 29 different variants (Figure 4; Supplementary Table 2). Two main functional categories were identified in the GWAS associations for these variants: immunological functions and neurological/behavioral traits. The associations with immune traits were expected given the results of previous balancing selection studies and the few well-characterized instances of LTBS. We identified many variants in LD with TSPs as associated with blood measurement phenotypes and diseases related to immune response failure (Supplementary Tables 2, 3, 4). Some of these traits include sarcoidosis (chromosome 4 near TSPs rs114383553 and rs17007061), ulcerative colitis and chronic inflammatory disease (chr2 near TSPs rs13426764/rs11694806), and Behçet's disease (chromosome 11 near TSPs rs2249268 and rs60427879).

We also found many neurological/behavioral related associations among TSP region variants. These traits include cognitive performance (chromosome 2 near rs13426764 and rs11694806 and chromosome 3 near rs9869178/rs2118072), chronotype (chromosome 4 near rs1887944 and rs7147645), addiction (alcohol use: chromosome 16 near rs9933768 and rs57790054; smoking: chromosome 2 near rs13426764 and rs11694806), risky behavior (automobile speeding propensity: chromosome 3:rs9869178/rs2118072), and experiencing mood swings (chromosome 2 near rs13426764 and rs11694806). In addition to the immune response and neurological categories, we observed associations in reproductive traits (polycystic ovary syndrome, testosterone levels), urate levels, pancreatic cancer, hepatocyte growth factor levels, and gut microbiota. We discuss several of these associations in more detail in following sections.

Phenome-wide association studies link TSPs to additional diverse traits

The growth of biobanks with linked genetic and phenotypic data has enabled the testing of the association of genetic variants with diverse traits within a single cohort. This PheWAS approach enables exploration of the functional and potentially pleiotropic effects of variants of interest (Bush et al., 2016). Using published associations from the UK Biobank, we analyzed the association of TSPs with 778 traits; all 125 of the TSP regions were tested. Overall, 21 TSPs in 16 regions had at least one genome-wide significant association ($P < 1E-8$, Figure 4; Supplementary Table 3). Though testing different phenotypes than the GWAS, these associations were qualitatively similar to the GWAS results, in that blood and immune system phenotypes had many associations with TSPs, but the TSPs were also associated with a more diverse set of phenotypes. We found associations in three major categories: immune response traits, body and physical measurements, and neuropsychiatric traits related to addiction. We also observed

associations with many other phenotypes, for example: walking habits, heart disorders, renal and kidney problems, pleural plaques, transient ischemic attack, cancerous tumor, and rupture of synovium and tendon (Supplementary Tables 3 and 4).

Figure 4. Genome- and phenome-wide association studies link TSPs to diverse traits. Genome-wide significant (P < 1E-8) associations from the GWAS Catalog (yellow), a PheWAS over the UK Biobank from the geneAtlas (purple), and other studies summarized in the GWAS Atlas (green) (Watanabe et al., 2019). Each dot represents an association between a TSP region and a trait. Many immune-related traits are associated with TSPs, but there are also associations with a wider variety of phenotypes including osseous, neurological, and nervous system traits. Five extreme associations with immunological and blood traits (P < 1E-60) were truncated for this visualization. Since few TSPs themselves were directly tested in GWAS, we include GWAS Catalog associations with tag variants in high LD ($r^2 > 0.8$) with TSPs. All associations are listed in Supplementary Tables 2–4.

Evidence of protein-coding function for TSPs

The TSPs were originally filtered by Leffler et al. (2013) to be non-coding; however, three of the 263 TSPs are coding variants: two variants containing synonymous alleles in the genes dynein (*DNHD1*) and in a transcript of *DNHD1*'s paralog *DNAH9*, and one non-synonymous variant in *PKD1L2*. This discrepancy is due to changes in genomic annotations over time. For example, PKD1L2 is a calcium channel with potential roles in kidney function. This gene is a polymorphic pseudogene in humans, with the reference genome encoding the pseudogene form; this likely explains why this coding variant was not identified as coding in previous studies. With the LD-expanded set, a total of 9 regions had non-synonymous codon changes (19 variants, Supplementary Table 5). Some of the genes containing these variants include: Sperm Associated Antigen 16 (*SPAG16*), which codes for two proteins that associate with the axoneme of sperm cells; Hepatocyte Nuclear Factor 4 Gamma (*HNF4G*) codes for a receptor involved in DNA binding transcription activity; Leucine Rich Colipase Like 1 (*LRCOL1*) which has enzyme activation activity and is involved in digestion. All of the coding variants had CADD scores suggestive of deleterious effects (scaled C-score > 20). Further work is needed to determine if these coding variants in high LD with the TSPs influence selection.

Illustrative examples of diverse functions associated with TSP regions

Integrating the above data, we found 79 TSP regions with two or more lines of functional evidence. This included twelve regions with annotations from at least four sources. To illustrate the diverse functions associated with TSPs, we highlight four of these regions. All the example regions were also detected by at least one of two other methods of finding genomic regions under balancing selection (Supplementary Table 7) (Bitarello et al., 2018; Siewert & Voight, 2020).

Body mass and alcohol intake. A TSP (rs57790054) on 16p12.3 (hg38.chr16:19992138-20043254) is strongly associated with several growth and body mass phenotypes as well as alcohol intake frequency (Figure 5A; P < 4E-10 for all). Another variant in high LD in Europeans (rs72771074, R^2=0.89) with a TSP (rs57790054) in this locus was associated with alcohol use disorder in a previous GWAS in a European cohort (P = 5E-8) (Sanchez-Roige et al., 2019). The nearest gene, *GPR139*, encodes for a G-protein coupled receptor expressed in the brain that is involved in alcohol drinking and withdrawal symptoms in rats (Kononoff et al., 2018). This region contains several variants in LD with TSPs in regulatory regions, such as CTCF binding sites (rs117293173, rs13338055, rs74011247, and rs79521770). One TSP (rs57790054) is an eQTL for the gene *KNOP1* (aka *C16orf88*), and 26 other LD SNPs are eQTL for both *KNOP1* and *IQCK*. These two genes have both been associated with obsessive compulsive disorder, among other diseases (Mattheisen et al., 2015). These results suggest that effects on growth and BMI or on addictive behaviors could be under LTBS. We note that there is some evidence of ethanol consumption in chimpanzees, but it is unclear how widespread its availability was over the past several million years (Hockings et al., 2015).

Urate levels. The TSPs (rs1839333, rs1913638) on 8q21.13 are both significantly associated (P < 2.0e-18) with uric acid levels in multiple GWAS in European and Asian ancestry populations (Figure 5B) (Kanai et al., 2018; Köttgen et al., 2013; Tin et al., 2019). These variants are also associated with a range of body mass traits in the UK Biobank. Another variant in this locus (rs2941471, R^2=0.97 and R^2=0.82 in East Asians and Europeans respectively) is associated with pancreatic cancer (p=7E-10). Though elevated uric acid in the blood is associated with many conditions, it is a marker for pancreatic cancer (Stotz et al., 2014). This locus also contains two non-synonymous variants in the gene *HNF4G*, a transcription factor expressed in the liver, kidney, and pancreas, in high LD with the TSPs (rs1805098 and rs2943549). The TSPs are also expression and splicing QTL for HNF4G (P = 1.9E-4 and 8.6E-6, respectively). Variants in *HNF4G* are associated with several traits, including the development of hyperuricemia (Chen et al., 2017).

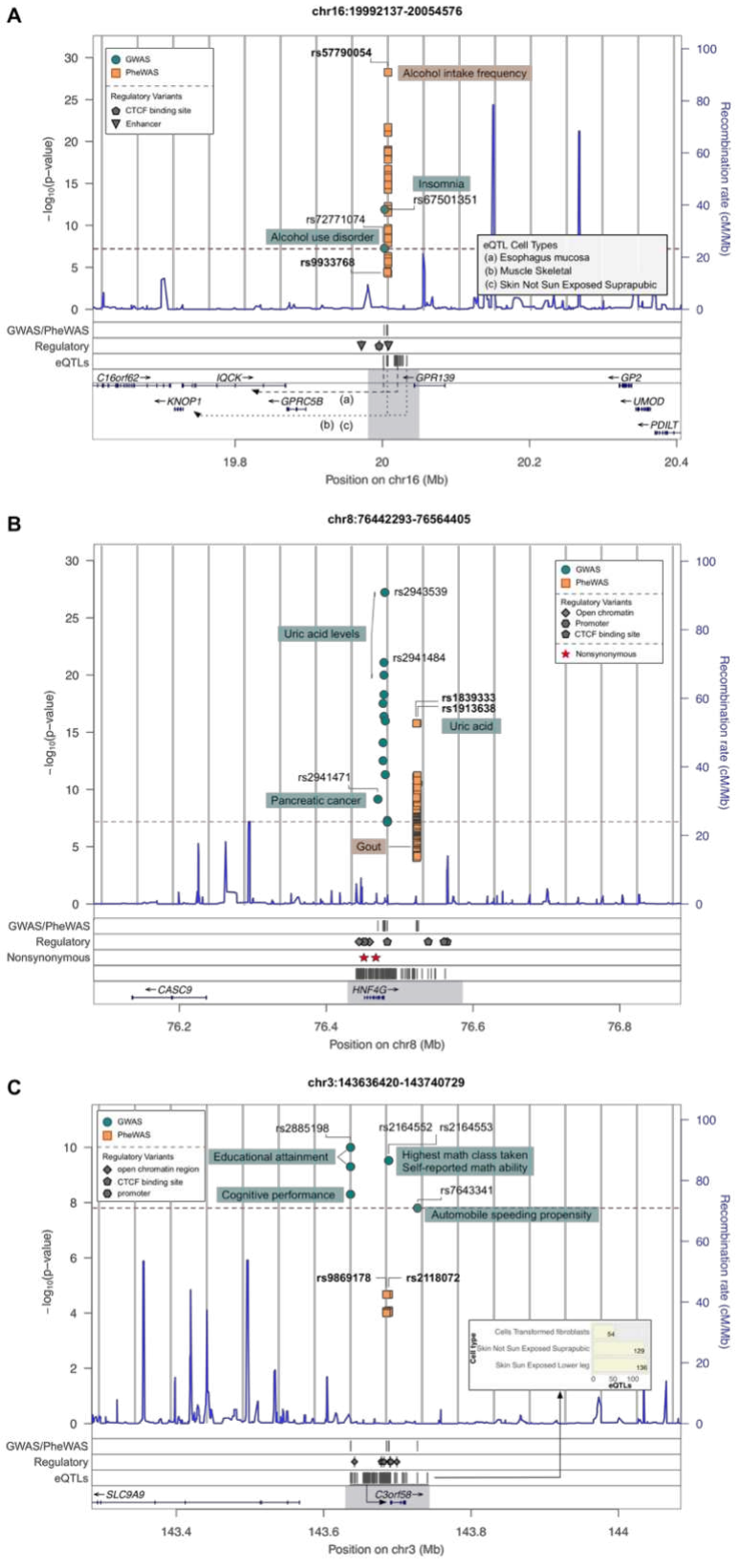

Figure 5. Illustrative examples of non-immune functions associated with TSPs.

(A) A TSP in 16p12.3 is associated with body mass index (BMI) and alcohol intake. Regional association plot showing statistically significant genome- and phenome-wide associations (threshold p ≤ 1E-08), regulatory annotations from Ensembl, and eQTLs from GTEx. One of the TSPs (rs57790054, orange) is associated with alcohol intake and several growth and body mass phenotypes in the UK Biobank. A variant in high LD (rs72771074, green) has been associated with alcohol use disorder in a previous GWAS. The TSP is also strongly associated with growth (comparative body size at age 10, 9.6e-21) and body mass index (3.5e-12). The TSPs are nearby *GPR139*, a gene encoding a G-protein coupled receptor expressed in the brain, whose expression levels influence alcohol drinking behavior in rats. The TSP region also contains several CTCF binding sites. TSP are shown in bold text.
(B) TSPs in 8q21.11 are associated with urate levels. Regional association plot showing statistically significant genome- and phenome-wide associations (P ≤ 1E-08), eQTL, regulatory and coding SNPs. LD SNPs in this region are associated with urate levels and pancreatic cancer. A TSP (rs1839333) is also associated with gout, although the p-value did not meet our strict threshold. TSP are shown in bold text. Figures created using LocusZoom (Pruim et al., 2011).
(C) TSP locus on 3q24. Regional association plot showing statistically significant genome- and phenome-wide associations (threshold p ≤ E-08), regulatory and eQTLs. This locus is characterized by neurological traits involved in educational attainment, cognitive performance, and risky behavior (automobile speeding propensity).

Chronotype. The TSPs (rs1887944, rs7147645) on 14q31.2 are both nominally associated (P <= 9.5e-6) with chronotype in a study of nearly 500,000 people (Jones et al., 2019). This is the only phenotype association with these variants. One variant in high LD (rs17119051) overlaps a regulatory region, but there are no eQTL among the TSPs or other LD variants. Asynchrony in chronotypes in a population has been proposed to be beneficial due to potential protection against predators and other dangers during the vulnerable hours of sleep. This so-called sentinel hypothesis developed out of the observation in Hadza hunter–gatherers of Tanzania that due to variation in sleep phase timing all members of a group are rarely asleep simultaneously (Samson et al., 2017). Thus, it is possible that LTBS maintained variation at this locus to promote chronotype diversity to mitigate the risks of sleeping.

Risky behavior and educational attainment. TSPs (rs9869178 and rs2118072) in 3q24 (hg19.chr3:143636420-143740729) are associated with a range of risky behaviors and educational attainment in both individual GWAS studies and the UK Biobank (Figure 5C). For example, they are associated with automobile speeding propensity (P = 4.4E-8) (Linnér, 2019) and with educational attainment (P = 2.4E-7) (Lee et al., 2018). The TSPs are also modestly associated with variation in brain white matter microstructure (Anterior corona radiata mean diusivities, P = 1.96E-6) (Zhao et al., 2019). Many of the variants in high LD with the TSPs in this region have similar associations and overlap annotated regulatory regions: regulatory open chromatin (rs10662845 and rs6766439), promoter (rs1898263, rs1992094, rs4431106,

rs7631704, rs7651567, and rs7653431), and CTCF binding sites (rs7628282, rs7650239, rs7650332, rs9840519, rs9840971, rs9878070, and *rs9840157*). Furthermore, the TSPs (and 159 high LD variants) are significant eQTLs ($P \leq 1E\text{-}5$) for the gene *DIPK2A* (*C3orf58*) across four GTEx tissues (small intestine terminal ileum, transformed fibroblasts, skin from the lower leg, and suprapubic skin). DIPK2A has not been comprehensively functionally characterized, but it contains a protein kinase domain and is broadly expressed, including in the developing and adult brain. Deletion of this gene has been linked to autism, and its expression is responsive to neuronal activity (Morrow et al., 2008). Associations with behavioral and cognitive traits must be interpreted with caution as these traits are very challenging to quantify and strongly influenced by social factors that may vary with other characteristics. Nonetheless, these associations point to an influence of the TSPs on behaviors relevant to risk tolerance. Thus, it is possible that maintaining a diversity of risk tolerance in human and chimpanzee populations has been beneficial.

Discussion

In this study we aimed to characterize the function of genomic regions with two or more TSPs in LD and close proximity. These variants have a deep ancestry in the common ancestor between humans and chimpanzees, and have persisted together in the genomes of both species for millions of years. Due to the maintenance of these polymorphisms over such long periods, they are likely evolving under LTBS. The majority of the non-coding TSP regions previously identified do not have known functions. To address this challenge, we identified functional annotations for 114 out of 125 TSP regions with the help of newly developed genomic annotation tools (Figure 2).

Our results suggest that non-coding TSPs likely maintained by LTBS have diverse functions beyond enabling a flexible immune response to pathogens. This expands on several recent studies of balancing selection over shorter timescales that have identified a small number of regions with functions outside the immune system (Bitarello et al., 2018; Sato & Kawata, 2018; Siewert & Voight, 2017; Viscardi et al., 2018).

The associations we identify suggest possible behavioral, neurological, and other traits that may have driven LTBS. In particular, our results provide support and candidate loci for previous hypotheses about the need for neurological and behavioral diversity in populations. For example, the chronotype association supports the sentinel hypothesis that variability in sleep patterns is the result of natural selection acting to reduce vulnerability of groups while members sleep at night (Samson et al., 2017). Similarly, selection has recently been shown to act on risk-taking behavior in anole lizards (Lapiedra et al., 2018). Thus, our identification of an association between a TSP and human risk-taking behavior (Figure 4C) suggests that LTBS may have maintained genetic variants that contribute to variation in risk taking behavior in humans and chimpanzees. A gene with evidence of eQTL in this region (*C3orf58*) encodes for a protein kinase and has been associated with autism and other neurological disorders (Dudkiewicz et al., 2013), adding further evidence of possible neurological drivers to LTBS.

Our results also raise the intriguing possibility that variants that modulate urate levels have been under LTBS. Uricase, the enzyme that metabolizes uric acid into an easily excreted water-soluble form in most mammals, has been lost in great apes. This gene was disabled by a series of mutations that slowly decreased activity over primate evolution, increasing the levels of uric acid in blood (Kratzer et al., 2014). It has been hypothesized that this loss of uricase activity was driven by increase fructose in primate diets due to fruit eating (Johnson et al., 2009; Kratzer et al., 2014). It has also been proposed that high levels of uric acid, a potent antioxidant, played an important role in the evolution of intelligence, acting as antioxidant in the brain (Álvarez-Lario & Macarrón-Vicente, 2010). However, as reflected in the associations with this locus, elevated uric acid levels contribute to many common diseases in modern humans, including chronic hypertension, cardiovascular disease, kidney and liver diseases, metabolic syndrome, diabetes, and obesity (Gustafsson & Unwin, 2013). This suggests potential functional tradeoffs at this locus. However, we emphasize that proving the environmental drivers of past selection is challenging.

Some of the phenotype associations we discovered may reflect manifestations of variation on traits in modern environments that could not be long-term drivers of balancing selection. As an extreme example, influence on smoking behavior could not have been the cause of LTBS given the relatively recent wide availability of nicotine. Though we note that there is some evidence of ethanol consumption in chimpanzees (Hockings et al., 2015). Even if they reflect modern environments, these associations provide hints about possible behavioral, neurological, or other traits that may have driven LTBS. For instance, plant chemicals can hijack reward systems in the brain that motivate repetition and learning (U.S. Department of Health & Human Services, 2016). The same systems that influence these action and consequently reproductive fitness potentially created a byproduct of excessive seeking of dopamine or other reward chemicals.

There are several caveats to our work. First, factors other than LTBS, such as high mutation rates and sequencing errors, can produce a TSP (Cheng & DeGiorgio, 2019; Gao et al., 2015). However, the presence of two or more TSPs in the regions we considered strongly suggest LTBS, and 80% of the regions had evidence of LTBS from an additional prediction method. Nonetheless, candidate regions of interest for future study should be further analyzed for possible confounders. Second, even with recent growth of genetic and phenotypic databases, our knowledge of the functions of most regions of the genome is sparse. Thus, failure to observe a functional association does not imply that a region does not have an important function. Third, the genome- and phenome-wide association tools we used are limited to the samples that have been analyzed; available data do not represent the full scope of human variation. Most of the individuals analyzed in available genetic association studies are of European ancestry. Variant functions and the ability to detect associations vary across human populations; however, we anticipate that TSPs should have functional effects across populations, unless modern environments have masked the pressure driving LTBS. Fourth, even in PheWAS, a limited number of phenotypes have been quantified across individuals, and these studies are focused on

a subset of clinically relevant rather evolutionarily relevant traits. Fifth, in some analyses, we considered annotations based on trait associations with variants in high LD ($r^2 > 0.8$) with TSPs. This could potentially introduce false positives if the variant also tags a different causal non-TSP variant. Given the long-term selection on TSPs, they are strong candidates for casual variants, but functional studies are needed to confirm these statistical associations. Finally, our analyses have focused on the human context; due to lack of functional data, it is not possible to explore the function of TSPs in chimpanzees. Nonetheless, we feel that our integration of genome-scale annotations and biobank data highlight the diversity of functions associated with LTBS.

Conclusions

In conclusion, we assign putative functions for TSP regions that likely persisted due to balancing selection dating back to at least the common ancestor of humans and chimpanzees. These annotations expand beyond immune functions to traits relevant to behavior, cognition, and body shape. Notably, we also find that most regions with multiple TSPs overlap gene regulatory annotations suggesting LTBS on gene expression levels. As methods improve for quantifying the effects of variants on gene regulation in different tissues and how these relate to organism-level phenotypes, we anticipate deeper mechanistic understanding of the functions and potential evolutionary pressures on these regions.

Methods

Trans-species polymorphisms

The initial set of 125 regions containing 263 TSPs analyzed in this study was published by Leffler et al. (2013). The set is composed of regions that: 1) contain at least two trans-species polymorphisms—i.e., variants that are segregating in both 51 Yoruba individuals in the 1000 Genomes Pilot 1 and 10 chimpanzees from the PanMap project—within 4 kb of each other in both species; and 2) are in high LD in humans and chimpanzees.

To increase our ability to identify annotations in each locus, the dataset was expanded to include variants in high LD (threshold R^2=0.8) with each of the TSPs as is common in association studies. We computed linkage disequilibrium with the TSP variants from 1000 Genomes Project Phase 3 data using rAggr a web tool developed by the University of Southern California (http://raggr.usc.edu). We considered LD in African, East Asia, and European populations. Variants with no reported RSID were excluded from the analysis. The dataset was thus expanded by 9,996 SNPs in high LD with the TSPs for a total of 10,259 SNPs.

LTBS prediction comparison

We compared the TSP regions in this study with LTBS candidate regions from two different methods developed to detect long-term balancing selection. BetaScan2 (Siewert & Voight, 2020) is a recent statistic for detecting balancing selection based enrichment for variants in a region with low variation in allele frequency and a deficit of substitutions. We identified overlaps

between the TSP regions and genomic regions detected by BetaScan2. Among the TSPs with Beta scores, 56% (57/102) had values greater than the 2.0 standardized beta score threshold used by the authors. Considering the entire TSP regions and LD, 83% (92/111) had at least one variant with a score above this threshold. We also computed overlap with regions identified by the NCD statistic (Bitarello et al., 2018). The overlap with the regions detected by NCD is 28% (35/125 regions). In total, 80% (100/125) of the TSP regions were supported by either the BetaScan2 or NCD.

Genome- and Phenome-wide associations

The GWAS Catalog collects variant-trait associations from published genome-wide association studies. The database is currently composed of more than 200,000 associations. We used the GWAS Catalog version October 2020 to find functional associations for the LTBS variants. The search was done using the BEDTools intersect function between the GWAS catalog and the LD-expanded TSP dataset (Quinlan, 2014).

PheWAS is an analysis strategy built on top of medical records with information about patient phenotypes and associated variants. The geneAtlas catalog (http://geneatlas.roslin.ed.ac.uk/phewas/) takes advantage of the data provided by the UK Biobank cohort, which contains medically relevant data from nearly 500,000 British individuals of European ancestry. This database contains 3 million variants in 778 traits. We matched our set of variants against the geneAtlas database to search for traits associated with LTBS.

GTEx eQTL data

To evaluate potential gene regulatory effects of TSPs in non-coding regions, we analyzed data from GTEx, a project developed to quantify the consequence of genetic variation on expression at the tissue level (https://www.gtexportal.org/home/). The GTEx project v8 data have identified eQTL across 50 tissues based on analyses of nearly 1,000 individuals to identify differential expression through SNP variation. The intersection between the TSPs and LD SNPs and the GTEx eQTL returned a large collection of TSP eQTL.

Other functional categories

We used ANNOVAR to identify variants overlapping protein coding regions. We used the Ensembl Regulatory Build (Zerbino et al., 2015) to identify variants overlapping regions with regulatory function. In all, 58 TSPs in 40 regions and 1334 LD SNPs in 114 regions overlapped regulatory annotations.

Enrichment for overlap with regulatory regions

We used a permutation framework to calculate whether TSPs were more enriched for overlap with regulatory regions than expected by chance (Benton et al., 2019). We quantified the number of overlapping TSPs for each type of regulatory region (open chromatin, promoter, enhancer, promoter-flanking, CTCF binding site, TF binding site). We then compared the observed TSP

overlap to a null distribution of expected overlap generated by randomly shuffling the regulatory regions 1000 times across the genome. We maintain the original length and chromosome distributions for shuffled regions and exclude all ENCODE blacklist and gap regions (Kundaje, 2013). We then computed an empirical p-value for the observed TSP overlap based on the distribution of overlaps for the set of matched shuffled regions.

Declarations

Availability of Data and Materials

The data underlying this article are available in the article and in its online supplementary material.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Institutes of Health [grant R35GM127087; grant T32LM012412], and the Burroughs-Wellcome Fund. The funders did not play any role in the study design, collection, analysis and interpretation of data, or in writing the manuscript.

Author's contributions

Conceptualization: JAC; Methodology: KV, MLB, JAC; Investigation: KV, MLB, JAC; Writing – Original Draft: KV, JAC; Writing – Review & Editing: KV, MLB, JAC; Funding Acquisition: JAC; Resources: JAC; Supervision: JAC.

Abbreviations

LTBS: Long-term balancing selection
TSP: Trans-species polymorphisms
LD: Linkage disequilibrium
eQTL: Expression quantitative trait loci
GTEx: Genotype-Tissue Expression
GWAS: Genome-wide association study
PheWAS: Phenome-wide association study

Acknowledgements

We thank Evonne McArthur, David Rinker, and other members of the Capra Lab for helpful comments on this work. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN.

REFERENCES

Álvarez-Lario, B., & Macarrón-Vicente, J. (2010). Uric acid and evolution. *Rheumatology*, *49*(11), 2010–2015. https://doi.org/10.1093/rheumatology/keq204

Azevedo, L., Serrano, C., Amorim, A., & Cooper, D. N. (2015). Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. *Human Genomics*, *9*(1). https://doi.org/10.1186/s40246-015-0043-1

Battivelli, E., Migraine, J., Lecossier, D., Yeni, P., Clavel, F., & Hance, A. J. (2011). Gag Cytotoxic T Lymphocyte Escape Mutations Can Increase Sensitivity of HIV-1 to Human TRIM5 , Linking Intrinsic and Acquired Immunity. *Journal of Virology*, *85*(22), 11846–11854. https://doi.org/10.1128/jvi.05201-11

Benton, M. L., Talipineni, S. C., Kostka, D., & Capra, J. A. (2019). Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *BMC Genomics*, *20*(1), 1–22. https://doi.org/10.1186/s12864-019-5779-x

Bitarello, B. D., De Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andres, A. M. (2018). Signatures of long-term balancing selection in human genomes. *Genome Biology and Evolution*, *10*(3), 939–955. https://doi.org/10.1093/gbe/evy054

Bush, W. S., Oetjens, M. T., & Crawford, D. C. (2016). Unravelling the human genome-phenome relationship using phenome-wide association studies. In *Nature Reviews Genetics* (Vol. 17, Issue 3, pp. 129–145). Nature Publishing Group. https://doi.org/10.1038/nrg.2015.36

Cagliani, R., Fumagalli, M., Biasin, M., Piacentini, L., Riva, S., Pozzoli, U., Bonaglia, M. C., Bresolin, N., Clerici, M., & Sironi, M. (2010). Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene. *Human Genetics*, *128*(6), 577–588. https://doi.org/10.1007/s00439-010-0884-6

Cagliani, R., Guerini, F. R., Fumagalli, M., Riva, S., Agliardi, C., Galimberti, D., Pozzoli, U., Goris, A., Dubois, B., Fenoglio, C., Forni, D., Sanna, S., Zara, I., Pitzalis, M., Zoledziewska, M., Cucca, F., Marini, F., Comi, G. P., Scarpini, E., … Sironi, M. (2012). A trans-specific polymorphism in ZC3HAV1 is maintained by long-standing balancing selection and may confer susceptibility to multiple sclerosis. *Molecular Biology and Evolution*, *29*(6), 1599–1613. https://doi.org/10.1093/molbev/mss002

Chen, B. D., Chen, X. C., Pan, S., Yang, Y. N., He, C. H., Liu, F., Ma, X., Gai, M. T., & Ma, Y. T. (2017). TT genotype of rs2941484 in the human HNF4G gene is associated with hyperuricemia in Chinese Han men. *Oncotarget*, *8*(16), 26918–26926. https://doi.org/10.18632/oncotarget.15851

Cheng, X., & DeGiorgio, M. (2019). Detection of Shared Balancing Selection in the Absence of Trans-Species Polymorphism. *Molecular Biology and Evolution*, *36*(1), 177–199. https://doi.org/10.1093/molbev/msy202

De Filippo, C., Key, F. M., Ghirotto, S., Benazzo, A., Meneu, J. R., Weihmann, A., Parra, G., Green, E. D., & Andrés, A. M. (2016). Recent Selection Changes in Human Genes under Long-Term Balancing Selection. *Molecular Biology and Evolution*, *33*(6), 1435–1447. https://doi.org/10.1093/molbev/msw023

DeGiorgio, M., Lohmueller, K. E., & Nielsen, R. (2014). A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLoS Genetics*, *10*(8). https://doi.org/10.1371/journal.pgen.1004561

Dudkiewicz, M., Lenart, A., & Pawłowski, K. (2013). A Novel Predicted Calcium-Regulated Kinase Family Implicated in Neurological Disorders. *PLoS ONE*, *8*(6).

https://doi.org/10.1371/journal.pone.0066427
Ganser-Pornillos, B. K., & Pornillos, O. (2019). Restriction of HIV-1 and other retroviruses by TRIM5. In *Nature Reviews Microbiology* (Vol. 17, Issue 9, pp. 546–556). Nature Publishing Group. https://doi.org/10.1038/s41579-019-0225-2
Gao, Z., Przeworski, M., & Sella, G. (2015). Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution*, *69*(2). https://doi.org/10.1111/evo.12567
Gustafsson, D., & Unwin, R. (2013). The pathophysiology of hyperuricaemia and its possible relationship to cardiovascular disease, morbidity and mortality. In *BMC Nephrology* (Vol. 14, Issue 1). https://doi.org/10.1186/1471-2369-14-164
Hockings, K. J., Bryson-Morrison, N., Carvalho, S., Fujisawa, M., Humle, T., McGrew, W. C., Nakamura, M., Ohashi, G., Yamanashi, Y., Yamakoshi, G., & Matsuzawa, T. (2015). Tools to tipple: Ethanol ingestion by wild chimpanzees using leaf-sponges. *Royal Society Open Science*, *2*(6). https://doi.org/10.1098/rsos.150150
Johnson, R. J., Sautin, Y. Y., Oliver, W. J., Roncal, C., Mu, W., Gabriela Sanchez-Lozada, L., Rodriguez-Iturbe, B., Nakagawa, T., & Benner, S. A. (2009). Lessons from comparative physiology: Could uric acid represent a physiologic alarm signal gone awry in western society? In *Journal of Comparative Physiology B: Biochemical, Systemic, and Environmental Physiology* (Vol. 179, Issue 1, pp. 67–76). Springer Verlag. https://doi.org/10.1007/s00360-008-0291-7
Jones, S. E., Lane, J. M., Wood, A. R., van Hees, V. T., Tyrrell, J., Beaumont, R. N., Jeffries, A. R., Dashti, H. S., Hillsdon, M., Ruth, K. S., Tuke, M. A., Yaghootkar, H., Sharp, S. A., Jie, Y., Thompson, W. D., Harrison, J. W., Dawes, A., Byrne, E. M., Tiemeier, H., … Weedon, M. N. (2019). Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. *Nature Communications*, *10*(343). https://doi.org/10.1038/s41467-018-08259-7
Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., Kubo, M., Okada, Y., & Kamatani, Y. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics*, *50*(3), 390–400. https://doi.org/10.1038/s41588-018-0047-6
Key, F. M., Teixeira, J. C., de Filippo, C., & Andrés, A. M. (2014). Advantageous diversity maintained by balancing selection in humans. *Current Opinion in Genetics and Development*, *29*, 45–51. https://doi.org/10.1016/j.gde.2014.08.001
Kononoff, J., Kallupi, M., Kimbrough, A., Conlisk, D., de Guglielmo, G., & George, O. (2018). Systemic and intra-habenular activation of the orphan G protein-coupled receptor GPR139 decreases compulsive-like alcohol drinking and hyperalgesia in alcohol-dependent rats. *ENeuro*, *5*(3). https://doi.org/10.1523/ENEURO.0153-18.2018
Köttgen, A., Albrecht, E., Teumer, A., Vitart, V., Krumsiek, J., Hundertmark, C., Pistis, G., Ruggiero, D., O'Seaghdha, C. M., Haller, T., Yang, Q., Tanaka, T., Johnson, A. D., Kutalik, Z., Smith, A. V., Shi, J., Struchalin, M., Middelberg, R. P. S., Brown, M. J., … Gieger, C. (2013). Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature Genetics*, *45*(2), 145–154. https://doi.org/10.1038/ng.2500
Kratzer, J. T., Lanaspa, M. A., Murphy, M. N., Cicerchi, C., Graves, C. L., Tipton, P. A., Ortlund, E. A., Johnson, R. J., & Gaucher, E. A. (2014). Evolutionary history and metabolic insights of ancient mammalian uricases. *Proceedings of the National Academy of Sciences*

667 *of the United States of America*, *111*(10), 3763–3768.
668 https://doi.org/10.1073/pnas.1320393111
669 Kundaje, A. (2013). *A comprehensive collection of signal artifact blacklist regions in the human*
670 *genome*.
671 ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/b
672 yFreeze/jan2011/blacklists/hg19-blacklist-README.pdf,
673 https://sites.google.com/site/anshulkundaje/projects/blacklists
674 Lapiedra, O., Schoener, T. W., Leal, M., Losos, J. B., & Kolbe, J. J. (2018). Predator-driven
675 natural selection on risk-taking behavior in anole lizards. *Science*, *360*(6392), 1017–1020.
676 https://doi.org/10.1126/science.aap9289
677 Lawlor, D. A., Ward, F. E., Ennis, P. D., Jackson, A. P., & Parham, P. (1988). HLA-A and B
678 polymorphisms predate the divergence of humans and chimpanzees. *Nature*, *335*(6187),
679 268–271. https://doi.org/10.1038/335268a0
680 Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A.,
681 Bowers, P., Sidorenko, J., Karlsson Linnér, R., Fontana, M. A., Kundu, T., Lee, C., Li, H.,
682 Li, R., Royer, R., Timshel, P. N., Walters, R. K., Willoughby, E. A., … Turley, P. (2018).
683 Gene discovery and polygenic prediction from a genome-wide association study of
684 educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), 1112–1121.
685 https://doi.org/10.1038/s41588-018-0147-3
686 Leffler, E. M., Gao, Z., Pfeifer, S., Ségurel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R.,
687 Wall, J. D., Sella, G., Donnelly, P., McVean, G., & Przeworski, M. (2013). Multiple
688 instances of ancient balancing selection shared between humans and chimpanzees. *Science*,
689 *340*(6127), 1578–1582. https://doi.org/10.1126/science.1234070
690 Linnér, R. K. (2019). Genome-wide association analyses of risk tolerance and risky behaviors in
691 over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature*
692 *Genetics*, *51*(2), 245–257. https://doi.org/10.1038/s41588-018-0309-3
693 Mao, R., Nie, H., Cai, D., Zhang, J., Liu, H., Yan, R., Cuconati, A., Block, T. M., Guo, J. T., &
694 Guo, H. (2013). Inhibition of Hepatitis B Virus Replication by the Host Zinc Finger
695 Antiviral Protein. *PLoS Pathogens*, *9*(7). https://doi.org/10.1371/journal.ppat.1003494
696 Mattheisen, M., Samuels, J. F., Wang, Y., Greenberg, B. D., Fyer, A. J., Mccracken, J. T.,
697 Geller, D. A., Murphy, D. L., Knowles, J. A., Grados, M. A., Riddle, M. A., Rasmussen, S.
698 A., Mclaughlin, N. C., Nurmi, E. L., Askland, K. D., Qin, H. D., Cullen, B. A., Piacentini,
699 J., Pauls, D. L., … Nestadt, G. (2015). Genome-wide association study in obsessive-
700 compulsive disorder: Results from the OCGAS. *Molecular Psychiatry*, *20*(3).
701 https://doi.org/10.1038/mp.2014.43
702 Mayer, W. E., Jonker, M., Klein, D., Ivanyi, P., van Seventer, G., & Klein, J. (1988). Nucleotide
703 sequences of chimpanzee MHC class I alleles: evidence for trans-species mode of
704 evolution. *The EMBO Journal*, *7*(9), 2765–2774. https://doi.org/10.1002/j.1460-
705 2075.1988.tb03131.x
706 Morrow, E. M., Yoo, S. Y., Flavell, S. W., Kim, T. K., Lin, Y., Hill, R. S., Mukaddes, N. M.,
707 Balkhy, S., Gascon, G., Hashmi, A., Al-Saad, S., Ware, J., Joseph, R. M., Greenblatt, R.,
708 Gleason, D., Ertelt, J. A., Apse, K. A., Bodell, A., Partlow, J. N., … Walsh, C. A. (2008).
709 Identifying autism loci and genes by tracing recent shared ancestry. *Science*, *321*(5886),
710 218–223. https://doi.org/10.1126/science.1157657
711 Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M.,
712 Abecasis, G. R., Willer, C. J., & Frishman, D. (2011). LocusZoom: Regional visualization

of genome-wide association scan results. *Bioinformatics*, *27*(13), 2336–2337. https://doi.org/10.1093/bioinformatics/btq419

Quinlan, A. R. (2014). BEDTools: The Swiss-Army tool for genome feature analysis. *Current Protocols in Bioinformatics*, *47*(1), 11–12. https://doi.org/10.1002/0471250953.bi1112s47

Samson, D. R., Crittenden, A. N., Mabulla, I. A., Mabulla, A. Z. P., & Nunn, C. L. (2017). Chronotype variation drives night-time sentinel-like behaviour in hunter–gatherers. *Proceedings of the Royal Society B: Biological Sciences*, *284*(20170967). https://doi.org/10.1098/rspb.2017.0967

Sanchez-Roige, S., Palmer, A. A., Fontanillas, P., Elson, S. L., Adams, M. J., Howard, D. M., Edenberg, H. J., Davies, G., Crist, R. C., Deary, I. J., McIntosh, A. M., Clarke, T. K., Elson, L., Fontanillas, P., Furlotte, N. A., Hinds, D. A., Huber, K. E., Kleinman, A., Litterman, N. K., … Wilson, C. H. (2019). Genome-wide association study meta-analysis of the alcohol use disorders identification test (AUDIT) in two population-based cohorts. *American Journal of Psychiatry*, *176*(2), 107–118. https://doi.org/10.1176/appi.ajp.2018.18040369

Sato, D. X., & Kawata, M. (2018). Positive and balancing selection on SLC18A1 gene associated with psychiatric disorders and human-unique personality traits. *Evolution Letters*, *2*(5), 499–510. https://doi.org/10.1002/evl3.81

Ségurel, L., Thompson, E. E., Flutre, T., Lovstad, J., Venkat, A., Margulis, S. W., Moyse, J., Ross, S., Gamble, K., Sella, G., Ober, C., & Przeworski, M. (2012). The ABO blood group is a trans-species polymorphism in primates. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(45), 18493–18498. https://doi.org/10.1073/pnas.1210603109

Siewert, K. M., & Voight, B. F. (2017). Detecting long-term balancing selection using allele frequency correlation. *Molecular Biology and Evolution*, *34*(11), 2996–3005. https://doi.org/10.1093/molbev/msx209

Siewert, K. M., & Voight, B. F. (2020). BetaScan2: Standardized Statistics to Detect Balancing Selection Utilizing Substitution Data. *Genome Biology and Evolution*, *12*(2), 3873–3877. https://doi.org/10.1093/gbe/evaa013

Stotz, M., Szkandera, J., Seidel, J., Stojakovic, T., Samonigg, H., Reitz, D., Gary, T., Kornprat, P., Schaberl-Moser, R., Hoefler, G., Gerger, A., & Pichler, M. (2014). Evaluation of uric acid as a prognostic blood-based marker in a large cohort of pancreatic cancer patients. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0104730

Teixeira, J. C., De Filippo, C., Weihmann, A., Meneu, J. R., Racimo, F., Dannemann, M., Nickel, B., Fischer, A., Halbwax, M., Andre, C., Atencia, R., Meyer, M., Parra, G., Pääbo, S., & Andrés, A. M. (2015). Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees, and bonobos. *Molecular Biology and Evolution*, *32*(5), 1186–1196. https://doi.org/10.1093/molbev/msv007

Tin, A., Marten, J., Halperin Kuhns, V. L., Li, Y., Wuttke, M., Kirsten, H., Sieber, K. B., Qiu, C., Gorski, M., Yu, Z., Giri, A., Sveinbjornsson, G., Li, M., Chu, A. Y., Hoppmann, A., O'Connor, L. J., Prins, B., Nutile, T., Noce, D., … Köttgen, A. (2019). Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nature Genetics*, *51*(10), 1459–1474. https://doi.org/10.1038/s41588-019-0504-x

Todorova, T., Bock, F. J., & Chang, P. (2015). Poly(ADP-ribose) polymerase-13 and RNA regulation in immunity and cancer. In *Trends in Molecular Medicine* (Vol. 21, Issue 6, pp. 373–384). Elsevier Ltd. https://doi.org/10.1016/j.molmed.2015.03.002

U.S. Department of Health & Human Services. (2016). Chapter 2: The neurobiology of

substance use, misuse, and addiction. In *Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health*.

Viscardi, L. H., Paixão-Côrtes, V. R., Comas, D., Salzano, F. M., Rovaris, D., Bau, C. D., Amorim, C. E. G., & Bortolini, M. C. (2018). Searching for ancient balanced polymorphisms shared between Neanderthals and modern humans. *Genetics and Molecular Biology*, *41*(1), 67–81. https://doi.org/10.1590/1678-4685-gmb-2017-0308

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T. J. C., van der Sluis, S., Andreassen, O. A., Neale, B. M., & Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, *51*(9). https://doi.org/10.1038/s41588-019-0481-0

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., & Flicek, P. R. (2015). The Ensembl Regulatory Build. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0621-5

Zhao, B., Zhang, J., Ibrahim, J. G., Luo, T., Santelli, R. C., Li, Y., Li, T., Shan, Y., Zhu, Z., Zhou, F., Liao, H., Nichols, T. E., & Zhu, H. (2019). Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n = 17,706). *Molecular Psychiatry*. https://doi.org/10.1038/s41380-019-0569-z

Figures

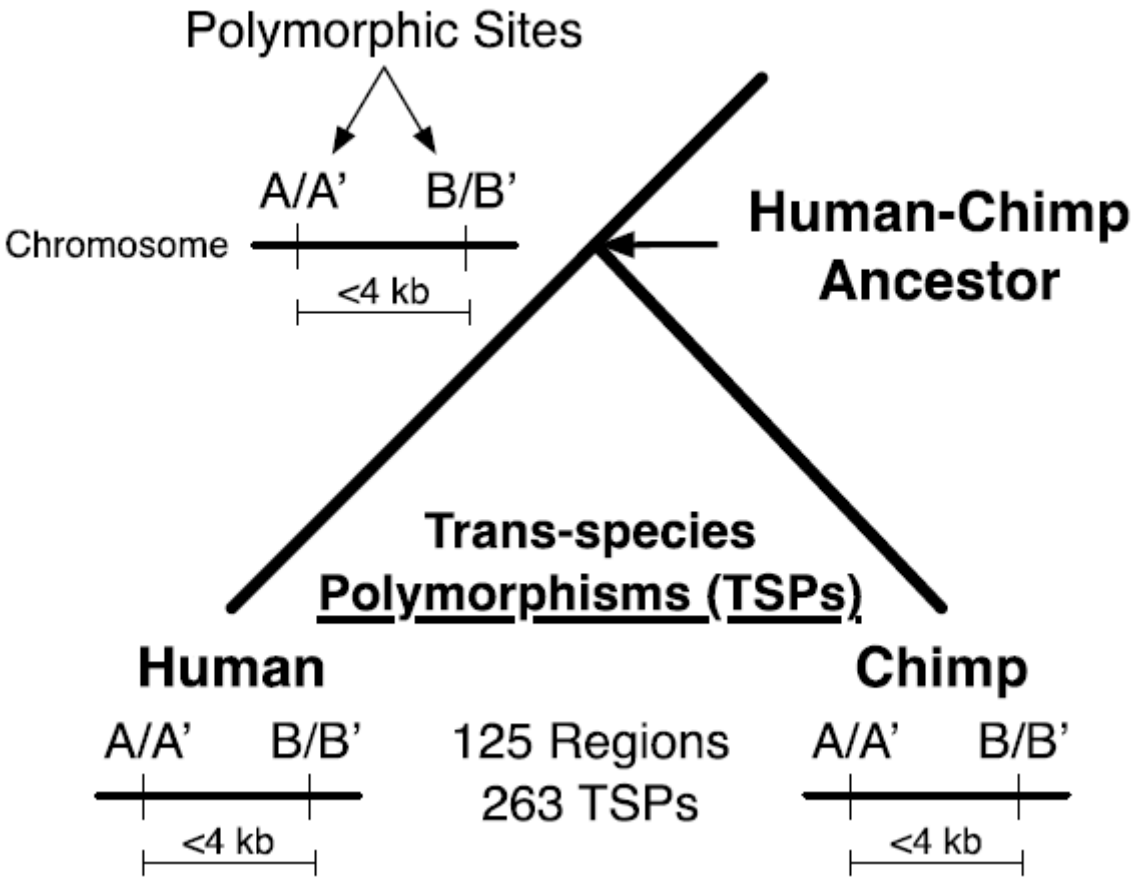

Figure 1

Trans-species polymorphisms (TSPs) likely resulting from long-term balancing selection (LTBS). Schematic showing the criteria used by Leffler et al. (2013) to identify TSPs likely maintained by LTBS. Each line represents a chromosome with polymorphisms segregating in a population. A/A' are two alleles segregating in both humans and chimpanzees at one site (i.e., a TSP), and B/B' are alleles segregating in both species at a nearby site. TSPs are very unlikely to appear nearby (within 4 kb) without the action of balancing selection. We consider 125 TSP regions containing 263 TSPs. Within these regions, multiple functional scenarios are possible. For example, one TSP may be under LTBS while the other is neutral, but maintained due to tight linkage. Alternatively, the TSPs may have epistatic functions and both be under selection. In addition to the 263 TSPs, we also considered functional associations with 9,996 variants in high LD ($r^2 > 0.8$) with a TSP at least one population from the 1000 Genomes Project (Supplementary Figure 1).

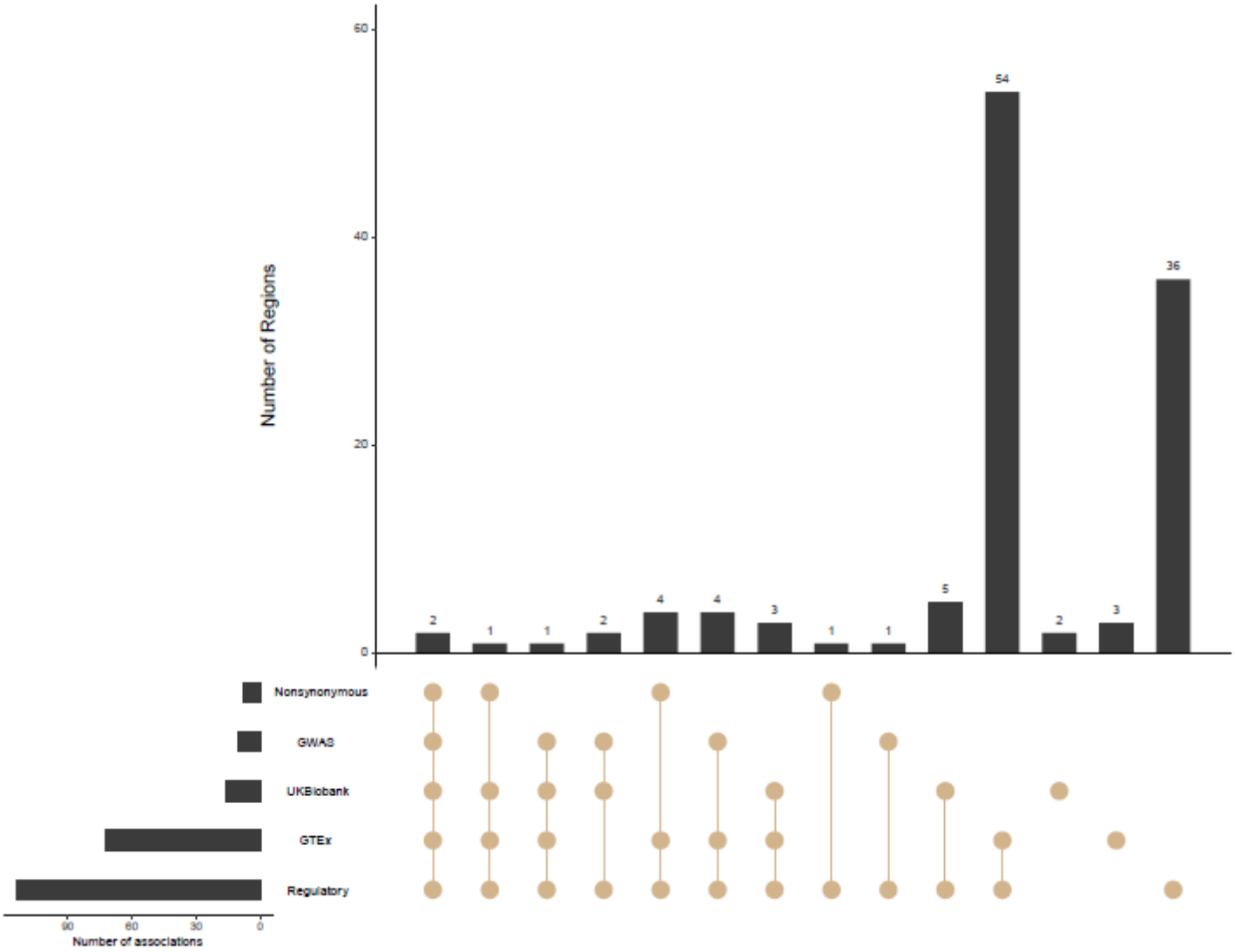

Figure 2

Functional annotations available for the expanded TSP regions. Summary of the annotations of each type available for TSP regions, including tagging variants in high LD with TSPs. A total of 119 out of 125 TSP regions contain at least one line of functional evidence. Multiple lines (two or more) are available for 78 regions.

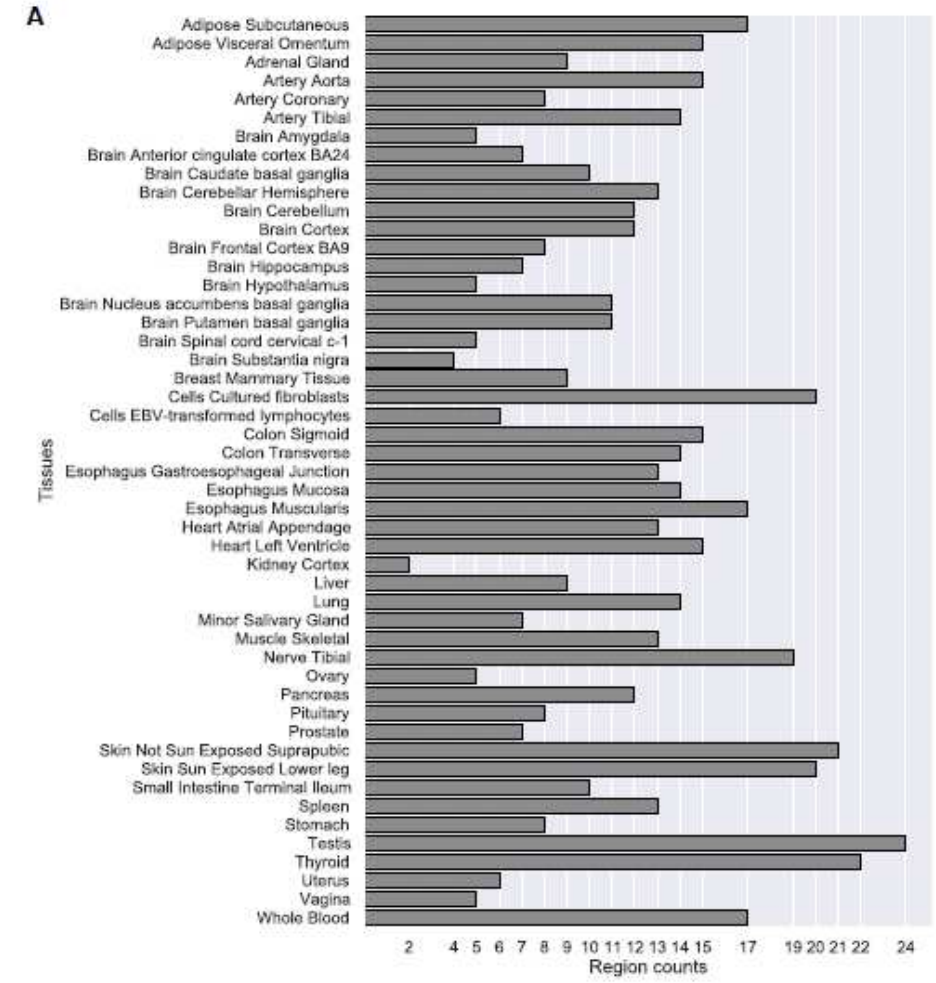

Figure 3

TSPs are eQTLs in diverse tissues and are depleted for overlap with promoters and CTCF sites. (A) The number of TSP regions that contain an eQTL for each GTEx tissue. Variation in TSP regions associates with gene expression in diverse tissues. The associated genes also have Gene Ontology (GO) annotations from diverse functional categories (Supplementary Figure 2). (B) TSP regions are significantly depleted

for overlap with promoters and CTCF sites compared to length- and chromosome-matched non-coding regions from the genomic background. The error bars represent 95% confidence intervals.

Figure 4

Genome- and phenome-wide association studies link TSPs to diverse traits. Genome-wide significant (P < 1E-8) associations from the GWAS Catalog (yellow), a PheWAS over the UK Biobank from the geneAtlas (purple), and other studies summarized in the GWAS Atlas (green) (Watanabe et al., 2019). Each dot represents an association between a TSP region and a trait. Many immune-related traits are associated with TSPs, but there are also associations with a wider variety of phenotypes including osseous, neurological, and nervous system traits. Five extreme associations with immunological and blood traits (P < 1E-60) were truncated for this visualization. Since few TSPs themselves were directly tested in GWAS, we include GWAS Catalog associations with tag variants in high LD (r2 > 0.8) with TSPs. All associations are listed in Supplementary Tables 2–4.

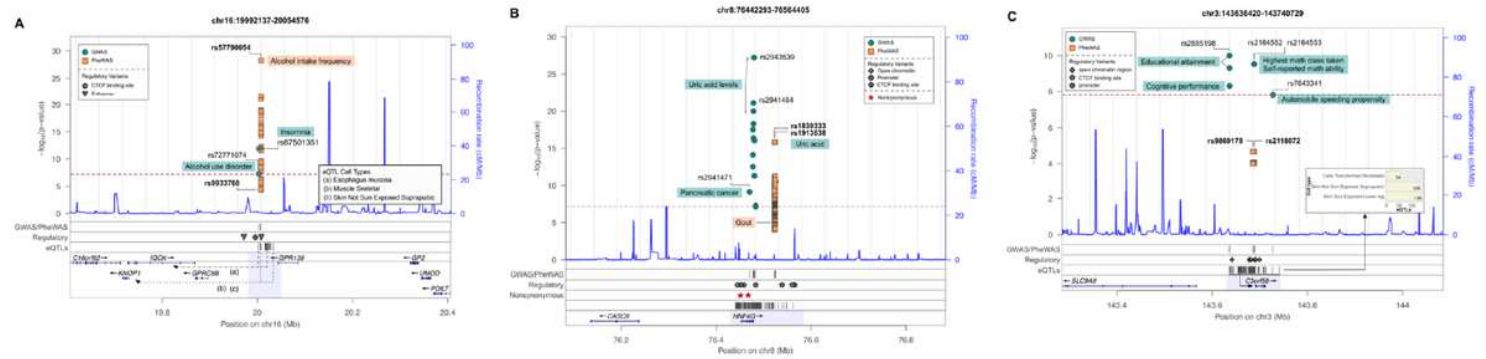

Figure 5

Illustrative examples of non-immune functions associated with TSPs. (A) A TSP in 16p12.3 is associated with body mass index (BMI) and alcohol intake. Regional association plot showing statistically significant genome- and phenome-wide associations (threshold p ≤ 1E-08), regulatory annotations from Ensembl, and eQTLs from GTEx. One of the TSPs (rs57790054, orange) is associated with alcohol intake and several growth and body mass phenotypes in the UK Biobank. A variant in high LD (rs72771074, green) has been associated with alcohol use disorder in a previous GWAS. The TSP is also strongly associated with growth (comparative body size at age 10, 9.6e-21) and body mass index (3.5e-12). The TSPs are nearby GPR139, a gene encoding a G-protein coupled receptor expressed in the brain, whose expression levels influence alcohol drinking behavior in rats. The TSP region also contains several CTCF binding sites. TSP are shown in bold text. (B) TSPs in 8q21.11 are associated with urate levels. Regional association plot showing statistically significant genome- and phenome-wide associations (P ≤ 1E-08), eQTL, regulatory and coding SNPs. LD SNPs in this region are associated with urate levels and pancreatic cancer. A TSP (rs1839333) is also associated with gout, although the p-value did not meet our strict threshold. TSP are shown in bold text. Figures created using LocusZoom (Pruim et al., 2011). (C) TSP locus on 3q24. Regional association plot showing statistically significant genome- and phenome-wide associations (threshold p ≤ E-08), regulatory and eQTLs. This locus is characterized by neurological traits involved in educational attainment, cognitive performance, and risky behavior (automobile speeding propensity).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplement.pdf
- supplementarytables.xlsx