Population-specific sequence and expression differentiation in Europeans

Xueyuan Jiang[1] and Raquel Assis*[,1,2]

[1] Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802

[2] Department of Biology, Pennsylvania State University, University Park, PA 16802

*Corresponding author: Raquel Assis

Department of Biology

Pennsylvania State University

208 Mueller Laboratory

University Park, PA, 16802

(814) 865-5804

rassis@psu.edu

Abstract

Much of the enormous phenotypic variation observed across human populations is thought to have arisen from events experienced as our ancestors peopled different regions of the world. However, little is known about the genes involved in these population-specific adaptations. Here we explore this problem by simultaneously examining population-specific sequence and expression differentiation in four human populations. In particular, we design a branch-based statistic to estimate population-specific differentiation in four populations, and apply this statistic to single nucleotide polymorphism (SNP) and RNA-seq data from Italian, British, Finish, and Yoruban populations. As expected, genome-wide estimates of sequence and expression differentiation each independently recapitulate the known demographic history of these four human populations, highlighting the utility of our statistic for identifying genic targets of population-specific adaptations. Application of our statistic reveals that genes containing large copy number variations (CNVs) have elevated levels of population-specific sequence and expression differentiation, consistent with the hypothesis that gene turnover is a key reservoir of adaptive variation. Further, European genes displaying population-specific sequence and expression differentiation are enriched for functions related to epigenetic regulation, immunity, and reproduction. Together, our findings illustrate that population-specific sequence and expression differentiation in humans may preferentially target genes with CNVs and play important roles in a diversity of adaptive and disease-related phenotypes.

Introduction

Human phenotypes vary widely across the globe. In particular, geographically separated populations often differ in skin pigmentation (Loomis 1967), hair color (Rees 2003), tooth morphology (Scott and Turner 1997; Hanihara and Ishida 2005), surface area to body mass ratio

(Katzmarzyk and Leonard 1998), and predisposition to diseases (Frank 2004). Much of this phenotypic variation is thought to have arisen due to a diversity of selective pressures experienced as early humans peopled the world and encountered novel environments (Sabeti et al. 2002; Voight et al. 2006), food sources (Sabeti et al. 2002), and pathogens (Diamond 2002; Jobling et al. 2013). As a result, uncovering the genetic targets of phenotypic variation among human populations is critical both for understanding past human adaptations (Sabeti et al. 2002) and for advancing future biomedical research (Jorde et al. 2001; Akey et al. 2004).

Due to the abundance of whole-genome sequence and polymorphism data for many human populations (Cann et al. 2002; International HapMap 3 Consortium 2010; The 1000 Genome Projects Consortium 2015), much work in the past several years has focused on elucidating and understanding sequence differentiation that occurred during human evolution (Li et al. 2008; Pickrell et al. 2009; Field et al. 2016). A common summary statistic for estimating sequence distances between two populations is the fixation index, F_{ST} (Wright 1951), which has been used to infer human demographic history (Hinds et al. 2005; Holsinger and Weir 2009; Keinan et al. 2009; Patterson et al. 2012; The 1000 Genome Projects Consortium 2015) and identify loci that may be targets of natural selection (Bowcock et al. 1991; Akey et al. 2002; Bersaglieri et al. 2004). However, because F_{ST} is a pairwise metric, it cannot identify the directionality of sequence differentiation nor be used as sole evidence for natural selection (Yi et al. 2010). To address this issue, Yi et al. (Yi et al. 2010) developed the Population Branch Statistic (PBS), a summary statistic that utilizes pairwise F_{ST} values among three populations to quantify sequence differentiation along each branch of their corresponding three-population tree. Genes with large PBS values on one branch represent loci that underwent population-specific sequence differentiation consistent with positive selection (Yi et al. 2010). PBS has been applied to

corroborate previously established targets of selection, including genes associated with skin pigmentation (Lamason et al. 2005) and dietary fat sources (Mathias et al. 2012), as well as to identify novel candidates for high-altitude adaptation in Tibetans (Yi et al. 2010).

However, because natural selection acts on phenotypes, analysis of sequence data only enables assessment of its indirect effects. For this reason, it may be advantageous to study selection more directly by exploiting the recent availability of RNA-seq data for several human populations (Lappalainen et al. 2013). Specifically, phenotypic evolution is thought to often occur through modifications in gene expression (King and Wilson 1975; Wang et al. 1996; Wray et al. 2003; Carroll 2005; Carroll 2008; Raj et al. 2010). Thus, studying gene expression differentiation among human populations may increase power for identifying loci underlying population-specific phenotypic variation. Indeed, like genetic differentiation, gene expression levels vary considerably across human populations (Cheung et al. 2005; Stranger et al. 2007) and often reflect population structure (Brown et al. 2018). Moreover, human genes with large PBS values are enriched for expression quantitative trait loci (eQTLs; Quiver and Lachance 2018).

In the present study, we simultaneously explore population-specific sequence and expression differentiation in four human populations: the Toscani in Italia (TSI), British in England and Scotland (GBR), Finnish in Finland (FIN), and Yoruba in Nigeria (YRI). For these analyses, we employ single nucleotide polymorphism (SNP; The 1000 Genome Projects Consortium 2015) and RNA-seq (Lappalainen et al. 2013) data from each population. First, we use F_{ST} (Wright 1951) and its analogue for estimating quantitative trait differentiation, P_{ST} (Leinonen et al. 2006), to quantify and examine genome-wide patterns of sequence and expression differentiation in the four human populations. Next, we adapt the approach of PBS (Yi et al. 2010) to P_{ST}, as well as

extend its computation to a four-population tree, enabling us to estimate both sequence and expression differentiation in each of the four human populations. Last, we apply these branch-based statistics to study population-specific sequence and expression differentiation in Europeans and uncover candidate genes and functional modules underlying adaptation in TSI, GBR, and FIN populations.

Results

Genome-wide patterns of sequence and expression differentiation in four human populations

A first goal of our study was to estimate sequence and expression differentiation among TSI, GBR, FIN, and YRI populations. To address this problem, we used SNP data (The 1000 Genome Projects Consortium 2015) to calculate the F_{ST} (Wright 1951), and RNA-seq data (Lappalainen et al. 2013) to calculate the P_{ST} (Leinonen et al. 2006), of every gene between each pair of the four human populations. We calculated F_{ST} using Hudson's formula (Hudson et al. 1992) and computed the ratio of averages to minimize bias (Reynolds et al. 1983; Weir and Cockerham 1984; International HapMap 3 Consortium 2010; Bhatia et al. 2013; see Materials and Methods for details). Due to environmental effects on P_{ST}, we followed the approach of Leinonen et al. (2006) in calculating P_{ST} under two contrasting scenarios: one in which environmental and non-additive genetic effects account for half of the observed expression variation (h^2 = 0.5), and a second in which only additive genetic effects contribute to the observed expression variation (h^2 = 1; see Materials and Methods for details). Examinations of Pearson's linear (r) and Spearman's nonlinear (ρ) correlations revealed small but significantly positive relationships between F_{ST} and P_{ST} in TSI-FIN, TSI-YRI, GBR-YRI, and FIN-YRI population pairs (Tables S1-S2), consistent with previous observations that sequence and expression differentiation are weakly or moderately

associated (Makova and Li 2003; Nuzhdin et al. 2004; Sartor et al. 2006; Hunt et al. 2012; Assis and Bachtrog 2013; Assis and Bachtrog 2015).

To explore genome-wide patterns of sequence and expression differentiation among the four human populations, we independently used F_{ST} and P_{ST} to construct gene trees and then infer population trees supported by majorities of these gene trees (see Materials and Methods for details). Population trees obtained from F_{ST} and P_{ST} (h^2 = 0.5 and h^2 =1) have the same topology (Figure 1), indicating that there is consistency between relationships inferred from genome-wide patterns of sequence and expression differentiation despite their weak correlations with one another. Moreover, as expected due to increased noise in gene expression data (Raser and O'Shea 2005), gene trees obtained from P_{ST} (h^2 = 0.5 and h^2 =1) are more variable than those derived from F_{ST}. Further, the topology of the population trees recapitulates human demographic history, in that TSI and GBR populations are most closely related to one another, the FIN population is an outgroup to TSI and GBR, and the YRI population is an outgroup to all three European populations. These results mirror those from similar studies of F_{ST} (Hinds et al. 2005; Holsinger and Weir 2009; Keinan et al. 2009; Patterson et al. 2012; The 1000 Genome Projects Consortium 2015), as well as findings that gene expression data often display population structure comparable to that of sequence data (Cheung et al. 2005; Stranger et al. 2007; Brown et al. 2018).

Estimation of population-specific sequence and expression differentiation on a four-population tree

Next, we sought to quantify population-specific sequence and expression differentiation of genes in each of the four human populations. For a three-population tree, population-specific

sequence differentiation of a gene along each branch can be estimated with PBS (Yi et al. 2010) (Figure 2A), which applies Equation 11.20 in Felsenstein (Felenstein 2004) to F_{ST}. In particular, considering the unrooted three-population tree shown in Figure 2A, the PBS value of a particular gene in population W is estimated as $PBS_W = \frac{1}{2}(E_{W,X} + E_{W,Y} - E_{X,Y})$, where $E_{W,X}$, $E_{W,Y}$, and $E_{X,Y}$ denote log-transformed F_{ST} between populations W and X, W and Y, and X and Y, respectively (Yi et al. 2010; see Materials and Methods for details). In a recent study, Equation 11.20 in Felsenstein (Felenstein 2004) was also applied to expression distances between orthologous genes to estimate branch lengths corresponding to lineage-specific expression divergence on a three-species tree (Assis 2018). Analogously, by substituting P_{ST} for F_{ST} in the formula for PBS (Yi et al. 2010), we can obtain the PBS corresponding to gene expression differentiation in population W on the three-population tree. To distinguish between these two PBS in our study, we will refer to the calculation with F_{ST} as the "sequence PBS", and the calculation with P_{ST} as the "expression PBS".

To enable quantification of population-specific sequence and expression differentiation in four human populations, we further extended the computation of each PBS to a four-population tree (Figure 2B) via application of Equation 12.6 in Felsenstein (Felenstein 2004). Henceforth, we will denote PBS as PBS_3 when applied to a three-population tree, and as PBS_4 when applied to a four-population tree. In particular, considering the unrooted four-population tree depicted in Figure 2B, the PBS value of a particular gene in population A can be estimated as $PBS_{4,W} = \frac{1}{4}(2E_{W,X} + E_{W,Y} + E_{W,Z} - E_{X,Y} - E_{X,Z})$, where $E_{W,X}$, $E_{W,Y}$, $E_{X,Y}$, and $E_{X,Z}$ denote log-transformed F_{ST} or P_{ST} of the gene between populations W and X, W and Y, X and Y, and X and Z, respectively (see Materials and Methods for details). We used this formula to compute the sequence PBS_4 and expression PBS_4 score of each gene in TSI, GBR, FIN, and YRI populations (Tables S3-S5).

Population-specific sequence and expression differentiation of genes with copy number variations

Gene duplications and deletions are key contributors to human genetic diversity (Sudmant et al. 2015). Moreover, because they are large-scale mutation events that may impact gene dosage, duplications and deletions have been implicated in numerous human diseases (Sebat et al. 2004; Kumar et al. 2008; Sharp et al. 2008; Weiss et al. 2008), as well as in adaptive events in many diverse species (Kaessmann 2010; Chen et al. 2013). As a result, genes containing copy number variations (CNVs) are thought to be more frequently targeted by natural selection than those without CNVs (Freeman et al. 2006; Nguyen et al. 2006). Indeed, genes with CNVs often display signatures of adaptation (Sudmant et al. 2015), and fixation of duplications and deletions has been associated with natural selection in many species (Freeman et al. 2006; Nguyen et al. 2006; Han et al. 2009b; Jiang and Assis 2017). Therefore, we hypothesized that genes with CNVs would have larger sequence and expression PBS_4 values than genes without CNVs.

To test this hypothesis, we compared the distributions of maximum PBS_4 values of genes with and without known human CNVs larger than 50bp (Figure 3; MacDonald et al. 2013; see Materials and Methods for details). Consistent with our prediction, both sequence and expression PBS_4 values are significantly elevated in genes with CNVs (Figure 3; $P < 0.05$ for all pairwise comparisons, two-sample permutation tests; see Materials and Methods for details). Therefore, genes containing large CNVs appear to undergo increased population-specific sequence and expression differentiation, supporting the hypothesis that both their sequences and expression patterns may be targeted by natural selection.

Relationships of population-specific sequence and expression differentiation to gene function in Europeans

A natural question that emerges from this analysis is what types of functional modules underlie population-specific sequence and expression differentiation in human populations. In addressing this question, it was important to exclude YRI, as it is an outgroup to the three European populations and therefore contains greater overall population-specific sequence and expression differentiation that is difficult to polarize. Hence, we only considered genes with large PBS_4 values in TSI, GBR, and FIN populations for these analyses. To globally assess functional modules contributing to population-specific sequence and expression differentiation in these European populations, we utilized annotation data from the Gene Ontology (GO) Consortium (Ashburner et al. 2000; GO Consortium 2018). In particular, GO terms classify genes by their molecular functions, cellular components, and biological processes (Ashburner et al. 2000; GO Consortium 2018). Therefore, to study relationships of population-specific sequence and expression differentiation to gene functions, we ranked genes by their sequence and expression PBS_4 values in each population, performed GO enrichment analysis on ranked lists, and extracted significantly overrepresented GO terms (Tables S6-S14; see Materials and Methods for details).

First, we examined GO enrichment for sequence PBS_4 values in TSI, GBR, and FIN populations (Tables S6-S8). Though the three populations do not share any GO terms, they each show enrichment in terms related to neuronal processes. In TSI, the most enriched GO term is "histone methylation", a process whereby methyl groups are attached to histone proteins. Several other enriched terms are also related to methylation, which is interesting because methylation is often used as a signal for gene activation or silencing (Jones 2012), suggesting

that population-specific sequence differentiation is related to epigenetic processes in TSI. Moreover, a few enriched GO terms in TSI are related to dynamics of cell membranes, which are common targets of adaptation (Hamblin and Di Rienzo 2000; Sabeti et al. 2006). In GBR, the most enriched GO term for sequence PBS_4 values is "branching involved in labyrinthine layer morphogenesis", a process whereby the branches of fetal placental villi are generated and organized. Therefore, increased population-specific sequence differentiation in GBR may be related to reproduction and fetal development. This is particularly intriguing because, though natural selection frequently targets reproductive phenotypes, often they are male-specific (Pröschel et al. 2006; Ellegren and Parsch 2007; Assis and Bachtrog 2013; Harrison et al. 2015). Yet this GO term specifically relates to female-specific functions, indicating that population-specific adaptation may be associated with female reproductive phenotypes in GBR. In FIN, the most enriched GO term for sequence PBS_4 values is "ncRNA transcription", a process whereby noncoding DNA is transcribed into RNA. Though noncoding RNAs are not translated into proteins, they often play important roles in gene regulation and expression (He and Hannon 2004; Mercer et al. 2009). Also interesting is that many other enriched GO terms in FIN are related to reproduction, and specifically to male gamete generation and spermatogenesis, which are often targets of adaptation (Li et al. 2002; Zhou and Bachtrog 2012).

Next, we examined GO enrichment for expression PBS_4 values in TSI, GBR, and FIN populations (Tables S9-S14). Using P_{ST} with $h^2 = 0.5$ and with $h^2 = 1$ yielded similar results, consistent with our expectations based on previous comparisons (see Figures 1 and 2). However, a surprising finding was the abundance of enriched GO terms for expression PBS_4 values relative to those for sequence PBS_4 values. Perhaps as a result, many enriched GO terms are shared by the three populations: "cell surface receptor signaling pathway", "developmental process", "positive

regulation of response to stimulus", "regulation of cell communication", "regulation of immune system process", "regulation of multicellular organismal process", and "signal transduction". Though most of these GO terms are quite general and difficult to interpret, it appears that population-specific expression differentiation in Europeans is often related to processes involved in signal transduction, immunity, reproduction, and development. Moreover, many related GO terms are enriched in Europeans, including those with the most enrichment in each population. The abundance of these GO terms is not surprising, particularly as such processes are frequent targets of natural selection (Barreiro and Quintana-Murci 2010; Fumagalli et al. 2011; Enard et al. 2016).

To glean further insight into specific genes potentially driving population-specific sequence and expression differentiation in Europeans, we performed literature searches on genes with the largest sequence and expression PBS_4 values in each population. The gene with the largest sequence PBS_4 value in both TSI and GBR is *MCM6,* or Minichromosome Maintenance Complex Component 6. *MCM6* is part of a protein complex essential for the initiation of eukaryotic genome replication (Labib et al. 2000). Moreover, two of its introns contain enhancers of its upstream gene *LCT*, or Lactase, one of which has a mutation that is prevalent in European populations and is thought to confer lactose tolerance in adulthood (Enattah et al. 2002; Troelsen et al. 2003). Consistent with this hypothesis, several genetic studies have identified *MCM6* and *LCT* as targets of recent positive selection in Europeans (Bersaglieri et al. 2004; Voight et al. 2006; Ranciaro et al. 2014; Cheng et al. 2017). Moreover, the gene with the second-largest sequence PBS_4 value in GBR is *LCT*. In contrast to the other two European populations, the gene with the largest sequence PBS_4 value in FIN is *NPIPA2*, or Nuclear Pore Complex Interacting Protein Family Member A2. *NPIPA2* is part of the nuclear pore complex of the cell

membrane, which regulates exchange between the nucleus and cytoplasm (Strambio-De-Castillia et al. 2010). Polymorphisms in *NPIPA2* are associated with several diverse classes of cancer (Dingerdissen et al. 2017), and its deletion is common in early-onset colorectal cancer (Perea et al. 2017). Moreover, both the localization of *NPIPA2* to the cell membrane and its function in transport across the membrane make it a likely target of natural selection (Tang and Presgraves 2009; Tracy et al. 2010).

The gene with the largest expression PBS_4 value (for P_{ST} with $h^2 = 0.5$ and $h^2 = 1$) in TSI is *PRRX1*, or Paired Related Homeobox 1. *PRRX1* is a DNA-associated protein that functions as a transcription coactivator and is involved in the establishment of diverse mesodermal muscle types during embryonic development (Martin et al. 1995). In particular, *PRRX1* is thought to play a critical role in craniofacial muscle development, as its variants are associated with several diseases that result in facial and neck malformations linked to sleep apnea (Martin et al. 1995; Urbizu et al. 2013). Further, *PRRX1* has also been associated with numerous cancers (Takahashi et al. 2013; Guo et al. 2015; Hirata et al. 2015; Jurecekova et al. 2016; Takano et al. 2016; Zhu et al. 2017), and is thought to mediate metastasis, or the migration and invasion of cancer cells into diverse tissues (Ocaña et al. 2012; Takahashi et al. 2013; Guo et al. 2015; Zhu et al. 2017). In GBR, the gene with the largest expression PBS_4 value (for P_{ST} with $h^2 = 0.5$ and $h^2 = 1$) is *PRKCB*, or Protein Kinase C Beta. *PRKCB* is involved in a diversity of cellular signaling pathway, including B cell activation during immune response (Lutzny et al. 2013), apoptosis (Reyland 2009), and autophagy (Patergnani et al. 2013). As a result, mutations in *PRKCB* are associated with numerous common diseases, including several cancers (Lutzny et al. 2013; Wallace et al. 2014; Antal et al. 2015) and autoimmune diseases (Han et al. 2009a; Sheng et al. 2010; Kawashima et al. 2017). The association with autoimmune diseases is particularly intriguing, as such genes are

often identified as targets of recent positive selection (Barreiro and Quintana-Murci 2010; Ramos et al. 2014). It is hypothesized that mutations that cause autoimmune response today may have provided pathogen resistance in the past (Barreiro and Quintana-Murci 2010). Last, in FIN, the gene with the highest expression PBS_4 value for P_{ST} with h^2 = 0.5 in FIN is *VDR*, or Vitamin D Receptor, whereas the gene with the highest expression PBS_4 value for P_{ST} with h^2 = 1 is *PLAC8,* or Placenta Specific 8. *VDR* interacts with vitamin D in the small intestine to facilitate calcium transportation into circulation (Holick 2006), and has been associated with vitamin D-dependent rickets (Holick 2006; Wagner and Greer 2008) and osteoporosis (Holick 2004). Skin exposure to solar ultraviolet radiation (UVR) produces about 90% of the vitamin D that an individual requires (Holick 2006), and living at high latitudes has been associated with vitamin D deficiency due to decreased UVR (Kimlin 2008; Chaplin and Jablonski 2009). Therefore, it is possible that expression differentiation of *VDR* may contribute to high latitude adaptation in FIN. *PLAC8* is also an interesting candidate, as it was first identified in human dendritic cells (Rissoan et al. 2002) and was later found to be expressed in interstitial extravillous trophoblast cells in the placenta, playing a key role in promoting their invasion and migration (Chang et al. 2018). *PLAC8* also facilitates the epithelial-to-mesenchymal transition (EMT) in colon cancer cells (Li et al. 2014) and has been associated with autoimmune diseases (Orrù et al. 2013). Therefore, expression differentiation of *PLAC8* in FIN may be related to its role in reproduction or disease.

Discussion

Identifying drivers of human phenotypic variation is crucial to understanding adaptive events that occurred in the past, as well as to developing population- and individual-targeted treatments for diseases in the future (Jorde et al. 2001; Sabeti et al. 2002; Akey et al. 2004). Though previous research (Sabeti et al. 2002; Akey et al. 2004; Voight et al. 2006) has made use

of abundant whole-genome and polymorphism data for many human populations (International HapMap 3 Consortium 2010; The 1000 Genome Projects Consortium 2015) to answer this question, simultaneously studying sequence and expression differentiation may provide unique insights into direct phenotypic targets of natural selection. In particular, it is thought that phenotypic evolution more often occurs through changes in gene regulation and expression, rather than their protein-coding sequences (King and Wilson 1975; Wang et al. 1996; Wray et al. 2003; Carroll 2005; Carroll 2008; Raj et al. 2010). For this reason, gene expression differentiation might better reflect phenotypic variation. Therefore, a major advantage of the present study is that we utilized from both sequence and expression data to address questions about population-specific differentiation in humans. Further, results from our combined analysis suggest that population-specific sequence and expression differentiation in humans may be attributed to several important biological processes, most notably immunity and reproduction, and also pinpoint many candidate genes for further study of human phenotypic variation in adaptation and disease.

A potential limitation of our study is the usage of RNA-seq data obtained from lymphoblastoid cell lines. In particular, the enrichment of immune-related functions in genes with high levels of population-specific expression differentiation may be attributed to usage of this cell line, rather than reflecting widespread evolutionary patterns of immunity genes across tissues. Yet it is important to note that associations between increased population-specific expression differentiation and immunity are consistent with previous findings. Specifically, immunity genes are among the fastest evolving genes in the human genome, likely due to adaptations to rapidly changing environments and introductions of novel pathogens (Barreiro and Quintana-Murci 2010; Fumagalli et al. 2011; Enard et al. 2016). Therefore, though observed patterns of

population-specific expression differentiation may not be representative of those in other cell types, genes with high population-specific expression differentiation should be further studied to examine their potential roles in human evolutionary history and disease. Regardless, future availability of RNA-seq data for multiple cell or tissue types in several populations will be invaluable for capturing complex patterns of population-specific expression differentiation and pinpointing genic targets of phenotypic variation among human populations.

Another caveat of these RNA-seq data (Lappalainen et al. 2013) is that TSI, GBR, and FIN are closely related European populations. Detection of population-specific differentiation is inherently difficult, as strong positive selection is thought to be rare in recent human evolution (Hernandez et al. 2011; The 1000 Genome Projects Consortium 2015). In addition, we expect sequence and expression differentiation to be correlated among these populations due to their shared ancestry and possible gene flow. Moreover, genetic and phenotypic differences among distantly related populations are better described than those among closely related populations, making it difficult to interpret our findings in the context of human phenotypes. Therefore, future availability of RNA-seq data from additional populations, particularly those that are more distantly related, will be critical to studying population-specific variation and its role in both human evolution and disease.

Despite the limitations of these data, a major advantage of our study is the design of PBS_4, a novel summary statistic that can be used to estimate population-specific sequence or expression differentiation in four populations. In particular, PBS_4 requires minimal assumptions about the data and can be used to rapidly estimate population-specific sequence or expression differentiation on a genome-wide scale. Further, because PBS_4 utilizes data from four

populations, branch lengths are more likely to represent true population-specific differentiation than differentiation that occurred ancestral to two populations, as is possible in a three-population scenario (Assis 2018). Therefore, though the dataset used in our study is not ideal in many respects, PBS_4 can easily be applied to existing or future datasets to quantify population-specific sequence or expression differentiation in humans and other species. In particular, we envision that application of PBS_4 to future human RNA-seq data from multiple cell lines or tissues and in many populations of varying divergence levels will shed light on complex questions about human evolutionary history and disease processes.

Materials and Methods

Population-genetic analyses

We downloaded the 1000 Genomes Project phase 3 dataset (The 1000 Genome Projects Consortium 2015) for TSI, GBR, FIN, and YRI populations from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/. After filtering out insertions, deletions, and single-nucleotide polymorphisms with Minor Allele Frequencies (MAF) less than 0.01, we were left with 15,145,123 SNPs. We calculated Hudson's F_{ST} for each SNP (Reynolds et al. 1983; Weir and Cockerham 1984; Bhatia et al. 2013). Then, we combined SNPs within the entire annotated region of each gene and computed the "ratio of averages" for Hudson's F_{ST} (Reynolds et al. 1983; Weir and Cockerham 1984; Bhatia et al. 2013). Because negative F_{ST} values are not defined (Wright 1951) and have no biological interpretation (Akey et al. 2002), we followed the standard of setting all negative F_{ST} values to zero (e.g., Nei 1990; Akey et al. 2002).

Gene expression analyses

We obtained RNA-seq data from lymphoblastoid cell lines in TSI, GBR, FIN, and YRI populations from the GEUVADIS project (Lappalainen et al. 2013). We excluded data from the population of Utah Residents with Northern and Western European Ancestry (CEU) because they were collected from an older cell line and display expression patterns that are inconsistent with their demographic history (Yuan et al. 2015). We quantified the abundance of transcripts in the remaining 371 individuals (93 in TSI, 94 in GBR, 95 in FIN, and 89 in YRI) using featureCount (Liao et al. 2013) with the long reads option (-L) and the GRCh37 human genome (Zerbino et al. 2017) as our reference. To normalize count data, we used the "median ratio method" method (Anders and Huber 2010) by implementing the estimateSizeFactors function in DESeq2 (Love et al. 2014). Next, we calculated the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) of each gene using DESeq2 (Love et al. 2014). We removed genes that contained fewer than 10 reads in each sample (lowly-expressed), were located on sex chromosomes, or were not protein-coding. For the remaining 13,075 genes, we log-transformed their FPKM values by $\log(\text{FPKM} + 1)$. We computed the P_{ST} for each gene as $P_{ST} = \frac{\sigma^2_{between}}{\sigma^2_{between} + 2h^2 \sigma^2_{within}}$ (Leinonen et al. 2006), where $\sigma^2_{between}$ is expression variance between populations, σ^2_{within} is expression variance within populations, and h^2 is heritability. For our analysis, we used $h^2 = 0.5$ and $h^2 = 1$ as was done previously (Leinonen et al. 2006), though we noted that the patterns in Figure 1 do not change as a function of h^2. When $h^2 = 1$, P_{ST} reduces to Q_{ST} (Spitze 1993), another common metric for differentiation of quantitative traits between populations.

Phylogenetic analyses

To infer population trees, we first built gene trees using the NEIGHBOR program in the PHYLIP package (Felsenstein 1993). We constructed gene trees using either F_{ST} or P_{ST} as input distances between populations. Application of the UPGMA algorithm in the NEIGHBOR program yielded

totals of 12,977 gene trees for F_{ST} and 13,075 gene trees for P_{ST}. Next, we used gene trees as input for the CONSENSE program in the PHYLIP package (Felsenstein 1993) and obtained rooted population trees supported by the majority of gene trees based on F_{ST} and P_{ST}. Specifically, the nodes in gene trees are included if they continue to resolve the population tree and do not contradict with more frequently occurring nodes. The number above each node in Figure 1 therefore represents its occurrence in all gene trees.

Calculation of PBS_4

We first computed the sequence or expression distance between populations as $E_{A,B} = -\log(1 - Z_{ST}(A,B))$, following the approach of Cavalli-Sforza (Cavalli-Sforza 1969), where Z_{ST} represents either F_{ST} or P_{ST} between populations A and B. We used these as input for calculations of sequence and expression PBS_4 values. Negative branch lengths were set to zero.

GO enrichment analyses

We performed all GO analyses on ranked lists of genes with the GOrilla tool (Eden et al. 2007; Eden et al. 2009). For each run, we output results for all enriched GO categories (process, function, and component) and set the P-value threshold to $P = 10^{-3}$.

Statistical analyses

All statistical analyses were performed in the R software environment (R Core Team 2013). Two-sample permutation tests were used to assess pairwise differences between all groups compared in Figure 3. For each test, we performed 1,000 permutations, using the difference between medians of groups as the test statistic. In particular, we computed the difference between the medians of the two groups for each permutation, and the P-value of the permutation test as the proportion of times the absolute value of this difference was greater

than or equal to the absolute value of the observed difference in the data. The significance of correlation coefficients shown in Table S1-2 were assessed via Student's *t* tests.

Acknowledgements

This work was supported by the National Science Foundation (DEB-1555981). Portions of this research were conducted with Advanced Cyber Infrastructure computational resources provided by the Institute for CyberScience at Pennsylvania State University (https://ics.psu.edu).

Disclosure declaration

The authors declare no conflict of interests.

References

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLOS Biol* 2: e286.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 12: 1805-1814.

Alves-Silva J, da Silva Santos M, Guimarães PE, Ferreira AC, Bandelt H-J, Pena SD, Prado VF. 2000. The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet* 67: 444-461.

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

Antal CE, Hudson AM, Kang E, Zanca C, Wirth C, Stephenson NL, Trotter EW, Gallegos LL, Miller CJ, Furnari FB. 2015. Cancer-associated protein kinase C mutations reveal kinase's role as tumor suppressor. *Cell* 160: 489-502.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT. 2000. Gene ontology: tool for the unification of biology. *Nature Genet* 25: 25-29.

Assis R. 2018. Lineage-specific expression divergence in grasses is associated with male reproduction, host-pathogen defense, and domestication. *Genome Biol Evol* 11: 207-219.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc Natl Acad Sci USA* 110: 17409-17414.

Assis R, Bachtrog D. 2015. Rapid divergence and diversification of mammalian duplicate gene functions. *BMC Evol Biol* 15: 138.

Baeuerle PA, Baltimore D. 1996. NF-κB: ten years after. *Cell* 87: 13-20.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet* 11: 17.

Basiri K, Belaya K, Liu WW, Maxwell S, Sedghi M, Beeson D. 2013. Clinical features in a large Iranian family with a limb-girdle congenital myasthenic syndrome due to a mutation in DPAGT1. *Neuromuscular Disord* 23: 469-472.

Belaya K, Finlayson S, Slater CR, Cossins J, Liu WW, Maxwell S, McGowan SJ, Maslau S, Twigg SR, Walls TJ. 2012. Mutations in DPAGT1 cause a limb-girdle congenital myasthenic syndrome with tubular aggregates. *Am J Hum Genet* 91: 193-201.

Bernard G, Chouery E, Putorti ML, Tétreault M, Takanohashi A, Carosso G, Clément I, Boespflug-Tanguy O, Rodriguez D, Delague V. 2011. Mutations of POLR3A encoding a catalytic subunit of RNA polymerase Pol III cause a recessive hypomyelinating leukodystrophy. *Am J Hum Genet* 89: 415-423.

1 Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE,
2 Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase
3 gene. *Am J Hum Genet* 74: 1111-1120.
4 Bhatia G, Patterson NJ, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST}: the
5 impact of rare variants. *Genome Res* 23:1514-1521.
6 Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL. 1991.
7 Drift, admixture, and selection in human evolution: a study with DNA polymorphisms.
8 *Proc Natl Acad Sci USA* 88: 839-843.
9 Brown BC, Bray NL, Pachter L. 2018. Expression reflects population structure. *PLOS Genet* 14:
10 e1007841.
11 Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. 2015. The genetic ancestry of african
12 americans, latinos, and european Americans across the United States. *Am J Hum Genet*
13 96: 37-53.
14 Burstein E, Hoberg JE, Wilkinson AS, Rumble JM, Csomos RA, Komarck CM, Maine GN, Wilkinson
15 JC, Mayo MW, Duckett CS. 2005. COMMD proteins, a novel family of structural and
16 functional homologs of MURR1. *J Biol Chem* 280: 22222-22232.
17 Cann HM, De Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-
18 Tamir B, Cambon-Thomsen A. 2002. A human genome diversity cell line panel. *Science*
19 296: 261-262.
20 Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. *Nature* 325:
21 31-36.
22 Canna SW, de Jesus AA, Gouni S, Brooks SR, Marrero B, Liu Y, DiMattia MA, Zaal KJ, Sanchez
23 GAM, Kim H. 2014. An activating NLRC4 inflammasome mutation causes

autoinflammation with recurrent macrophage activation syndrome. *Nature Genet* 46: 1140-1146.

Carrera IA, Matthijs G, Perez B, Cerdá CP. 2012. DPAGT1-CDG: Report of a patient with fetal hypokinesia phenotype. *American Journal of Medical Genetics Part A* 158: 2027-2030.

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLOS Biol* 3: e245.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134: 25-36.

Cavalli-Sforza LL. 1969. Human diversity. In *Proc 12th Int Congr Genet*, Vol 2, pp. 405-416.

Chang W-L, Liu Y-W, Dang Y-L, Jiang X-X, Xu H, Huang X, Wang Y-L, Wang H, Zhu C, Xue L-Q. 2018. PLAC8, a new marker for human interstitial extravillous trophoblast cells, promotes their invasion and migration. *Development* dev. 148932.

Chaplin G, Jablonski NG. 2009. Vitamin D and the evolution of human depigmentation. *Am J Phys Anthrop* 139: 451-461.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet* 14: 645-660.

Cheng X, Xu C, DeGiorgio M. 2017. Fast and robust detection of ancestral selective sweeps. *Mol Ecol* 26: 6871-6891.

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437: 1365-1369.

Chiu Y-H, MacMillan JB, Chen ZJ. 2009. RNA polymerase III detects cytosolic DNA and induces type I interferons through the RIG-I pathway. *Cell* 138: 576-591.

Cooper MD. 2015. The early history of B cells. *Nat Rev Immunol* 15: 191-197.

Corach D, Lao O, Bobillo C, van Der Gaag K, Zuniga S, Vermeulen M, Van Duijn K, Goedbloed M, Vallone PM, Parson W. 2010. Inferring continental ancestry of Argentineans from autosomal, Y-chromosomal and mitochondrial DNA. *Ann Hum Genet* 74: 65-76.

Devchand PR, Keller H, Peters JM, Vazquez M, Gonzalez FJ, Wahli W. 1996. The PPARα–leukotriene B4 pathway to inflammation control. *Nature* 384: 39.

Diamond J. 2002. Evolution, consequences and future of plant and animal domestication. *Nature* 418: 700-707.

Dingerdissen HM, Torcivia-Rodriguez J, Hu Y, Chang T-C, Mazumder R, Kahsay R. 2017. BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery. *Nucleic Acids Res* 46: D1128-D1136.

Dumay-Odelot H, Durrieu-Gaillard S, Da Silva D, Roeder RG, Teichmann M. 2010. Cell growth- and differentiation-dependent regulation of RNA polymerase III transcription. *Cell Cycle* 9: 3711-3723.

Eden E, Lipson D, Yogev S, Yakhini Z. 2007. Discovering motifs in ranked lists of DNA sequences. *PLOS Comput Biol* 3: e39.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10: 48.

Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet* 8: 689-698.

Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. *Elife* 5: e12469.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nature Genet* 30: 233-237.

Felsenstein J. 1993. *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein.

Felsenstein J. 2004. *Inferring phylogenies*. Sinauer associates Sunderland, MA.

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI. 2016. Detection of human adaptation during the past 2000 years. *Science* 354: 760-764.

Ford-Hutchinson A, Bray M, Doig MV, Shipley M, Smith M. 1980. Leukotriene B, a potent chemokinetic and aggregating substance released from polymorphonuclear leukocytes. *Nature* 286: 264.

Franca MM, Han X, Funari MF, Lerario AM, Nishi MY, Fontenele EG, Domenice S, Jorge AA, Garcia-Galiano D, Elias CF. 2019. Exome sequencing reveals POLR3H gene as a novel cause of Primary Ovarian Insufficiency. *J Clin Endocrinol Metab*.

Frank SA. 2004. Genetic predisposition to cancer—insights from population genetics. *Nat Rev Genet* 5: 764-772.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME. 2006. Copy number variation: new insights in genome diversity. *Genome Res* 16: 949-961.

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

GO Consortium. 2018. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* 47: D330-D338.

Guo J, Fu Z, Wei J, Lu W, Feng J, Zhang S. 2015. PRRX1 promotes epithelial–mesenchymal transition through the Wnt/β-catenin pathway in gastric cancer. *Med Oncol* 32: 393.

Hamblin MT, Di Rienzo A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 66: 1669-1679.

Han J-W, Zheng H-F, Cui Y, Sun L-D, Ye D-Q, Hu Z, Xu J-H, Cai Z-M, Huang W, Zhao G-P. 2009a. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nature Genet* 41: 1234-1237.

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009b. Adaptive evolution of young gene duplicates in mammals. *Genome Res* 19: 859-867.

Hanihara T, Ishida H. 2005. Metric dental variation of major human populations. *Am J Phys Anthrop* 128: 287-298.

Harrison PW, Wright AE, Zimmer F, Dean R, Montgomery SH, Pointer MA, Mank JE. 2015. Sexual selection drives evolution and rapid turnover of male gene expression. *Proc Natl Acad Sci USA* 112: 4393-4398.

Hayden MS, Ghosh S. 2004. Signaling to NF-κB. *Genes Del* 18: 2195-2224.

He L, Hannon GJ. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* 5: 522-531.

Heifetz A, Elbein AD. 1977. Solubilization and properties of mannose and N-acetylglucosamine transferases involved in formation of polyprenyl-sugar intermediates. *J Biol Chem* 252: 3057-3063.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331: 920-924.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* 307: 1072-1079.

Hirata H, Sugimachi K, Takahashi Y, Ueda M, Sakimura S, Uchi R, Kurashige J, Takano Y, Nanbara S, Komatsu H. 2015. Downregulation of PRRX1 confers cancer stem cell-like properties and predicts poor prognosis in hepatocellular carcinoma. *Ann Surg Oncol* 22: 1402-1409.

1 Holick MF. 2004. Vitamin D: importance in the prevention of cancers, type 1 diabetes, heart
2 disease, and osteoporosis. *Am J Clin Nutr* 79: 362-371.
3 Holick MF. 2006. Resurrection of vitamin D deficiency and rickets. *J Clin Invest* 116: 2062-2072.
4 Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining,
5 estimating and interpreting F_{ST}. *Nat Rev Genet* 10: 639-650.
6 Hu B, Elinav E, Huber S, Booth CJ, Strowig T, Jin C, Eisenbarth SC, Flavell RA. 2010. Inflammation-
7 induced tumorigenesis in the colon is regulated by caspase-1 and NLRC4. *Proc Natl Acad*
8 *Sci USA* 107: 21635-21640.
9 Hudson RR, Slatkin M, Maddison W. 1992. Estimation of levels of gene flow from DNA sequence
10 data. *Genetics* 132: 583-589.
11 Hunt BG, Ometto L, Keller L, Goodisman MA. 2012. Evolution at two levels in fire ants: the
12 relationship between patterns of gene expression and protein sequence evolution. *Mol*
13 *Biol Evol* 30: 263-271.
14 International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in
15 diverse human populations. *Nature* 467: 52-58.
16 Iqbal Z, Shahzad M, Vissers LE, Van Scherpenzeel M, Gilissen C, Razzaq A, Zahoor MY, Khan SN,
17 Kleefstra T, Veltman JA. 2013. A compound heterozygous mutation in DPAGT1 results in
18 a congenital disorder of glycosylation with a relatively mild phenotype. Eur J Hum Genet
19 21: 844-849.
20 Jiang X, Assis R. 2017. Natural selection drives rapid functional evolution of young Drosophila
21 duplicate genes. *Mol Biol Evol* 34:3089-3098.
22 Jobling M, Hurles M, Tyler-Smith C. 2013. *Human evolutionary genetics: origins, peoples &*
23 *disease*. Garland Science.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 13: 484-492.

Jorde L, Watkins WS, Bamshad M. 2001. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum Mol Genet* 10: 2199-2207.

Jurecekova J, Grendár M, Babušíková E, Kliment J, Dobrota D, Halašová E. 2016. Genome-wide association study of prostate cancer in population of Slovak men. *European Urology Supplements* 15: e1343-e1344.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* 20: 1313-1326.

Katzmarzyk PT, Leonard WR. 1998. Climatic influences on human body size and proportions: ecological adaptations and secular trends. *Am J Phys Anthrop* 106: 483-503.

Kawashima M, Hitomi Y, Aiba Y, Nishida N, Kojima K, Kawai Y, Nakamura H, Tanaka A, Zeniya M, Hashimoto E. 2017. Genome-wide association studies identify PRKCB as a novel genetic susceptibility locus for primary biliary cholangitis in the Japanese population. *Hum Mol Genet* 26: 650-659.

Keinan A, Mullikin JC, Patterson N, Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nature Genet* 41: 66-70.

Kimlin MG. 2008. Geographic location and vitamin D synthesis. *Mol Aspects Med* 29: 453-461.

King M-C, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188: 107-116.

Kumar A, Kumar J, Gadodia A, Chumber S, Aggarwal L. 2008. Multiple short-segment colonic duplications. *Pediatr Radiol* 38: 567-570.

Labib K, Tercero JA, Diffley JF. 2000. Uninterrupted MCM2-7 function required for DNA replication fork progression. *Science* 288: 1643-1647.

1 Lamason RL, Mohideen M-AP, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X,
2 Humphreville VR, Humbert JE. 2005. SLC24A5, a putative cation exchanger, affects
3 pigmentation in zebrafish and humans. *Science* 310: 1782-1786.
4 Lappalainen T, Sammeth M, Friedländer MR, AC't Hoen P, Monlong J, Rivas MA, Gonzalez-Porta
5 M, Kurbatova N, Griebel T, Ferreira PG. 2013. Transcriptome and genome sequencing
6 uncovers functional variation in humans. *Nature* 501: 506-511.
7 Lauberth SM, Nakayama T, Wu X, Ferris AL, Tang Z, Hughes SH, Roeder RG. 2013. H3K4me3
8 interactions with TAF3 regulate preinitiation complex assembly and selective gene
9 activation. *Cell* 152: 1021-1036.
10 Lehrman MA. 1991. Biosynthesis of N-acetylglucosamine-PP-dolichol, the committed step of
11 asparagine-linked oligosaccharide assembly. *Glycobiology* 1: 553-562.
12 Leinonen T, Cano J, Mäkinen H, Merilä J. 2006. Contrasting patterns of body shape and neutral
13 genetic divergence in marine and lake populations of threespine sticklebacks. *J Evol Biol*
14 19: 1803-1812.
15 Li C, Ma H, Wang Y, Cao Z, Graves-Deal R, Powell AE, Starchenko A, Ayers GD, Washington MK,
16 Kamath V. 2014. Excess PLAC8 promotes an unconventional ERK2-dependent EMT in
17 colon cancer. *J Clin Invest* 124: 2172-2187.
18 Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS,
19 Feldman M, Cavalli-Sforza LL. 2008. Worldwide human relationships inferred from
20 genome-wide patterns of variation. *Science* 319: 1100-1104.
21 Li W-H, Yi S, Makova K. 2002. Male-driven evolution. *Curr Opin Genet Dev* 12: 650-656.
22 Liang W, Ouyang S, Shaw N, Joachimiak A, Zhang R, Liu Z-J. 2011. Conversion of D-ribulose 5-
23 phosphate to D-xylulose 5-phosphate: new insights from structural and biochemical
24 studies on human RPE. *The FASEB Journal* 25: 497-504.

Liao Y, Smyth GK, Shi W. 2013. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30: 923-930.

Liu YF, Swart M, Ke Y, Ly K, McDonald FJ. 2013. Functional interaction of COMMD3 and COMMD9 with the epithelial sodium channel. *Am J Physiol Renal Physiol* 305: F80-F89.

Loomis WF. 1967. Skin-Pigment Regulation of Vitamin-D Biosynthesis in Man: Variation in solar ultraviolet at different latitudes may have caused racial differentiation in man. *Science* 157: 501-506.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550.

Lutzny G, Kocher T, Schmidt-Supprian M, Rudelius M, Klein-Hitpass L, Finch AJ, Dürig J, Wagner M, Haferlach C, Kohlmann A. 2013. Protein kinase c-β-dependent activation of NF-κB in stromal cells is indispensable for the survival of chronic lymphocytic leukemia B cells in vivo. *Cancer Cell* 23: 77-92.

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2013. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 42: D986-D992.

Makova KD, Li W-H. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res* 13: 1638-1645.

Malhotra JD, Kaufman RJ. 2011. ER stress and its functional link to mitochondria: role in cell survival and death. *Cold Spring Harb Perspect Biol* 3: a004424.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLOS Biol* 3: e357.

Martin JF, Bradley A, Olson EN. 1995. The paired-like homeo box gene MHox is required for early events of skeletogenesis in multiple lineages. *Genes Del* 9: 1237-1249.

Mathias RA, Fu W, Akey JM, Ainsworth HC, Torgerson DG, Ruczinski I, Sergeant S, Barnes KC, Chilton FH. 2012. Adaptive evolution of the FADS gene cluster within Africa. *PLOS ONE* 7: e44926.

McDonald MJ, Rice DP, Desai MM. 2016. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* 531: 233-236.

Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet* 10: 155-159.

Miao EA, Alpuche-Aranda CM, Dors M, Clark AE, Bader MW, Miller SI, Aderem A. 2006. Cytoplasmic flagellin activates caspase-1 and secretion of interleukin 1β via Ipaf. *Nature Immunol* 7: 569-575.

Miao EA, Mao DP, Yudkovsky N, Bonneau R, Lorang CG, Warren SE, Leaf IA, Aderem A. 2010. Innate immune detection of the type III secretion apparatus through the NLRC4 inflammasome. *Proc Natl Acad Sci USA* 107: 3076-3080.

Montenegro G, Rebelo AP, Connell J, Allison R, Babalini C, D'Aloia M, Montieri P, Schüle R, Ishiura H, Price J. 2012. Mutations in the ER-shaping protein reticulon 2 cause the axon-degenerative disorder hereditary spastic paraplegia type 12. *J Clin Invest* 122: 538-544.

Nguyen D-Q, Webber C, Ponting CP. 2006. Bias of selection on human copy-number variants. *PLOS Genet* 2: e20.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. 2004. Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Mol Biol Evol* 21: 1308-1317.

Ocaña OH, Córcoles R, Fabra Á, Moreno-Bueno G, Acloque H, Vega S, Barrallo-Gimeno A, Cano A, Nieto MA. 2012. Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. *Cancer cell* 22: 709-724.

1 Olivo PD, Van de Walle MJ, Laipis PJ, Hauswirth WW. 1983. Nucleotide sequence evidence for
2 rapid genotypic shifts in the bovine mitochondrial DNA D-loop. *Nature* 306: 400-402.
3 Orrù V, Steri M, Sole G, Sidore C, Virdis F, Dei M, Lai S, Zoledziewska M, Busonero F, Mulas A.
4 2013. Genetic variants regulating immune cell levels in health and disease. *Cell* 155: 242-
5 256.
6 Patergnani S, Marchi S, Rimessi A, Bonora M, Giorgi C, Mehta KD, Pinton P. 2013. PRKCB/protein
7 kinase C, beta and the mitochondrial axis as key regulators of autophagy. *Autophagy* 9:
8 1367-1385.
9 Patterson NJ, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich
10 D. 2012. Ancient admixture in human history. *Genetics* 192:1065-1093.
11 Perea J, García JL, Pérez J, Rueda D, Arriba M, Rodríguez Y, Urioste M, González-Sarmiento R.
12 2017. NOMO-1 gene is deleted in early-onset colorectal cancer. *Oncotarget* 8: 24429-
13 24436.
14 Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers
15 RM, Feldman MW. 2009. Signals of recent positive selection in a worldwide sample of
16 human populations. *Genome Res* 19: 826-837.
17 Pröschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of Drosophila genes with
18 sex-biased expression. *Genetics* 174: 893-900.
19 Quiver MH, Lachance J. 2018. Adaptive eQTLs reveal the evolutionary impacts of pleiotropy and
20 tissue-specificity, while contributing to health and disease in human populations.
21 *BioRxiv*: 444737
22 R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for
23 Statistical Computing, Vienna, Austria. 2013.

Raj A, Rifkin SA, Andersen E, Van Oudenaarden A. 2010. Variability in gene expression underlies incomplete penetrance. *Nature* 463: 913-918.

Ramos PS, Shaftman SR, Ward RC, Langefeld CD. 2014. Genes associated with SLE are targets of recent positive selection. *Autoimmune Dis* 2014.

Ranciaro A, Campbell MC, Hirbo JB, Ko W-Y, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA. 2014. Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am J Hum Genet* 94: 496-510.

Raser JM, O'Shea EK. 2005. Noise in gene expression: origins, consequences, and control. *Science* 309:2010-2013.

Rees JL. 2003. Genetics of hair and skin color. *Annals Rev Genet* 37: 67-90.

Reyland ME. 2009. Protein kinase C isoforms: multi-functional regulators of cell life and death. *Front Biosci (Landmark Ed)* 14: 2386-2399.

Reynolds J, Weir BS, Cockerham CC. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105: 767-779.

Rissoan M-C, Duhen T, Bridon J-M, Bendriss-Vermare N, Péronne C, de Saint Vis B, Briere F, Bates EE. 2002. Subtractive hybridization reveals the expression of immunoglobulinlike transcript 7, Eph-B1, granzyme B, and 3 novel transcripts in human plasmacytoid dendritic cells. *Blood* 100: 3295-3303.

Roebroek AJ, Contreras B, Pauli IG, Van de Ven WJ. 1998. cDNA Cloning, Genomic Organization, and Expression of the HumanRTN2Gene, a Member of a Gene Family Encoding Reticulons. *Genomics* 51: 98-106.

Romberg N, Al Moussawi K, Nelson-Williams C, Stiegler AL, Loring E, Choi M, Overton J, Meffre E, Khokha MK, Huttner AJ. 2014. Mutation of NLRC4 causes a syndrome of enterocolitis and autoinflammation. *Nature Genet* 46: 1135-1139.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832-837.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen T, Altshuler D, Lander E. 2006. Positive natural selection in the human lineage. *Science* 312: 1614-1620.

Sartor MA, Zorn AM, Schwanekamp JA, Halbleib D, Karyala S, Howell ML, Dean GE, Medvedovic M, Tomlinson CR. 2006. A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in Xenopus. *Nucleic Acids Res* 34: 185-200.

Scott GR, Turner CG. 1997. *Anthropology of modern human teeth*. Cambridge University Press Cambridge.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M. 2004. Large-scale copy number polymorphism in the human genome. *Science* 305: 525-528.

Selcen D, Shen X-M, Brengman J, Li Y, Stans AA, Wieben E, Engel AG. 2014. DPAGT1 myasthenia and myopathy: genetic, phenotypic, and expression studies. *Neurology* 82: 1822-1830.

Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, Schroer RJ, Novara F, De Gregori M, Ciccone R. 2008. A recurrent 15q13. 3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genet* 40: 322-328.

Sheng Y-J, Gao J-P, Li J, Han J-W, Xu Q, Hu W-L, Pan T-M, Cheng Y-L, Yu Z-Y, Ni C. 2010. Follow-up study identifies two novel susceptibility loci PRKCB and 8p11. 21 for systemic lupus erythematosus. *Rheumatology* 50: 682-688.

Spitze K. 1993. Population structure in Daphnia obtusa: quantitative genetic and allozymic variation. *Genetics* 135: 367-374.

Strambio-De-Castillia C, Niepel M, Rout MP. 2010. The nuclear pore complex: bridging nuclear transport and gene regulation. *Nat Rev Mol Cell Biol* 11: 490-501.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D. 2007. Population genomics of human gene expression. *Nature Genet* 39: 1217.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M. 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science* 349: aab3761.

Takahashi Y, Sawada G, Kurashige J, Uchi R, Matsumura T, Ueo H, Takano Y, Akiyoshi S, Eguchi H, Sudo T. 2013. Paired related homoeobox 1, a new EMT inducer, is involved in metastasis and poor prognosis in colorectal cancer. *Br J Cancer* 109: 307-311.

Takano S, Reichert M, Bakir B, Das KK, Nishida T, Miyazaki M, Heeg S, Collins MA, Marchand B, Hicks PD. 2016. Prrx1 isoform switching regulates pancreatic cancer invasion and metastatic colonization. *Genes Del* 30: 233-247.

Tang S, Presgraves DC. 2009. Evolution of the Drosophila nuclear pore complex results in multiple hybrid incompatibilities. *Science* 323: 779-782.

The 1000 Genomes Projects Consortium. 2015. A global reference for human genetic variation. *Nature* 526: 68-74.

Tracy C, Río J, Motiwale M, Christensen SM, Betrán E. 2010. Convergently recruited nuclear transport retrogenes are male biased in expression and evolving under positive selection in Drosophila. *Genetics* 184: 1067-1076.

Troelsen JT, Olsen J, Møller J, SjÖstrÖm H. 2003. An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology* 125: 1686-1694.

Urbizu A, Toma C, Poca MA, Sahuquillo J, Cuenca-Leon E, Cormand B, Macaya A. 2013. Chiari malformation type I: a case-control association study of 58 developmental genes. *PLOS ONE* 8: e57241.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLOS Biol* 4: e72.

Wagner CL, Greer FR. 2008. Prevention of rickets and vitamin D deficiency in infants, children, and adolescents. *Pediatrics* 122: 1142-1152.

Wallace JA, Pitarresi JR, Sharma N, Palettas M, Cuitiño MC, Sizemore ST, Yu L, Sanderlin A, Rosol TJ, Mehta KD. 2014. Protein kinase C Beta in the tumor microenvironment promotes mammary tumorigenesis. *Front Oncol* 4: 87.

Wang D, Marsh JL, Ayala FJ. 1996. Evolutionary changes in the expression pattern of a developmentally essential gene in three Drosophila species. *Proc Natl Acad Sci USA* 93: 7103-7107.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358-1370.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T. 2008. Association between microdeletion and microduplication at 16p11. 2 and autism. N Engl J Med 358: 667-675.

White RJ. 2005. RNA polymerases I and III, growth control and cancer. *Nat Rev Mol Cell Biol* 6: 69.

Wilson AC, Cann RL, Carr SM, George M, Gyllensten UB, Helm-Bychowski KM, Higuchi RG, Palumbi SR, Prager EM, Sage RD. 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. *Biol J Linn Soc Lond* 26: 375-400.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* 20: 1377-1419.

Wright S. 1951. The genetical structure of populations. Vol 15, pp. 323-354, *Annu. Eugenics*.

Wu X, Rush JS, Karaoglu D, Krasnewich D, Lubinsky MS, Waechter CJ, Gilmore R, Freeze HH. 2003. Deficiency of UDP-GlcNAc: dolichol phosphate N-acetylglucosamine-1 phosphate transferase (DPAGT1) causes a novel congenital disorder of glycosylation type Ij. *Hum Mutat* 22: 144-150.

Würde AE, Reunert J, Rust S, Hertzberg C, Haverkämper S, Nürnberg G, Nürnberg P, Lehle L, Rossi R, Marquardt T. 2012. Congenital disorder of glycosylation type Ij (CDG-Ij, DPAGT1-CDG): extending the clinical and molecular spectrum of a rare disease. *Mol Genet Metab* 105: 634-641.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329: 75-78.

Yuan Y, Tian L, Lu D, Xu S. 2015. Analysis of genome-wide RNA-sequencing data suggests age of the CEPH/Utah (CEU) lymphoblastoid cell lines systematically biases gene expression profiles. *Sci Rep* 5: 7960.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG. 2017. Ensembl 2018. *Nucleic Acids Res* 46: D754-D761.

Zhao Y, Yang J, Shi J, Gong Y-N, Lu Q, Xu H, Liu L, Shao F. 2011. The NLRC4 inflammasome receptors for bacterial flagellin and type III secretion apparatus. *Nature* 477: 596-600.

Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science* 337: 341-345.

Zhu H, Sun G, Dong J, Fei L. 2017. The role of PRRX1 in the apoptosis of A549 cells induced by cisplatin. *Am J Transl Res* 9: 396-402.

Zurek N, Sparks L, Voeltz G. 2011. Reticulon short hairpin transmembrane domains are used to shape ER tubules. *Traffic* 12: 28-41.

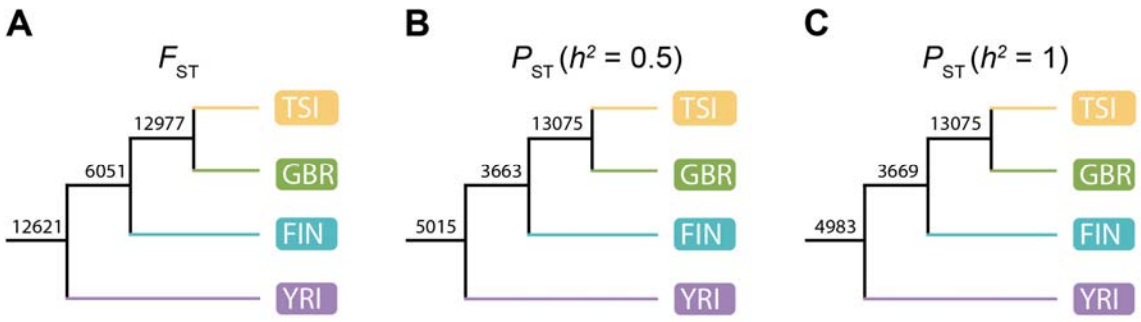

Figure 1. Relationships among TSI, GBR, FIN, and YRI populations inferred from genome-wide patterns of sequence and expression differentiation. Population trees supported by the majority of gene trees built using (*A*) F_{ST}, (*B*) P_{ST} with h^2 = 0.5, and (*C*) P_{ST} with h^2 = 1. Numbers indicate occurrences of corresponding nodes in all gene trees (see Materials and Methods for details).

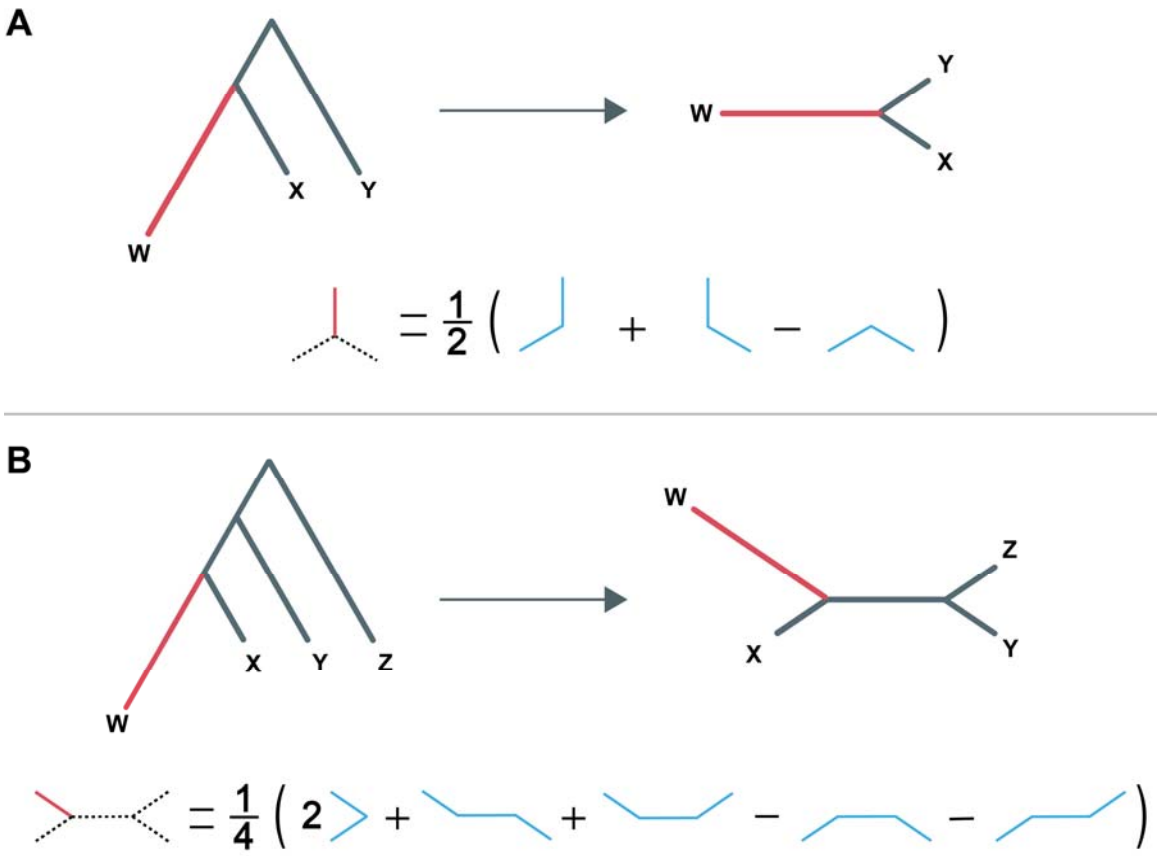

Figure 2. Schematic of PBS calculations. (*A*) Application of PBS_3 to population W in a tree relating populations W, X, and Y. The rooted three-population tree is unrooted (top), and the length of branch W is computed by applying the formula shown (bottom) to pairwise sequence (F_{ST}) or expression (P_{ST}) distances between populations. (*B*) Application of PBS_4 to population W in a tree relating populations W, X, Y, and Z. The rooted four-population tree is unrooted (top), and the length of branch W is computed by applying the formula shown (bottom) to pairwise sequence (F_{ST}) or expression (P_{ST}) distances between populations.

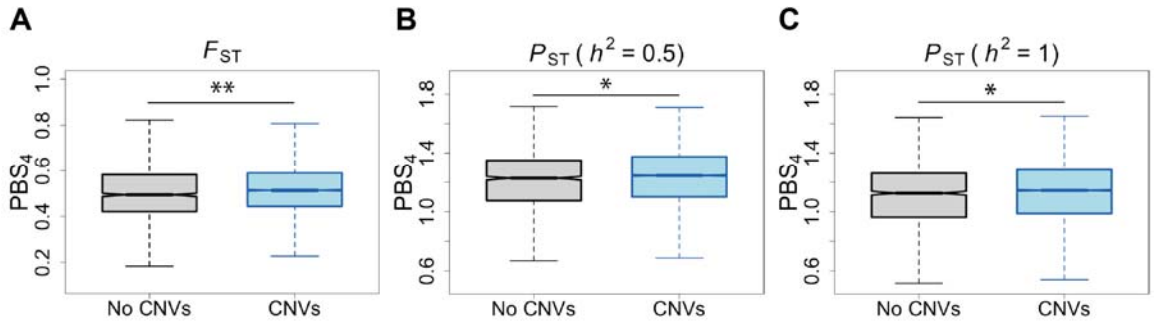

Figure 3. PBS_4 values of genes with and without CNVs. Distributions of cube root transformed (*A*) sequence PBS_4 values, (*B*) expression PBS_4 values with h^2 = 0.5, and (*C*) expression PBS_4 values with h^2 = 1 of genes with (blue) and without (gray) CNVs. * $P < 0.05$, ** $P < 0.001$. (See Materials and Methods for details).

1

2

3