Improved Characterization of Balancing Selection Genome Wide and a Detailed Look at HLA Genes

Tristan J. Hayeck[1,2], Timothy L. Mosbruger[1], Jonathan P Bradfield,[3] Adam G. Gleason[1], George Damianos[1], Jamie L. Duke[1], Laura K. Conlin, Tychele N. Turner[4], Mahdi Sarmady[1,2], Dimitri S. Monos[1,2]

[1]Division of Genomic Diagnostics, Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.

[2]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

[3]Quantinuum Research LLC, Philadelphia, PA, USA

[4]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA

Abstract

Balancing selection occurs when different evolutionary pressures impact the fitness of multiple alleles, resulting in increased allelic diversity in the population. A new statistical method was developed to test for selection, improving inference by using efficient Bayesian techniques to test for density and strength of linkage disequilibrium. Evolutionary simulation studies showed that the method consistently outperformed existing methods. Using this methodology, we tested for novel signals of balancing selection genome wide in 500 samples from phased trios. Several novel signals of selection appeared in CYP2A7, GPC6, and CNR2 across multiple ancestries. Additionally, tests in SIRPA demonstrate dramatically strong selection signal, significantly higher than previously observed. Well-known signals around olfactory genes and the MHC, containing HLA genes associated with the immune response, also demonstrated strong signatures of selection. So, utilizing data from the 17th IHIW, a follow up analysis was then performed by leveraging over seven thousand HLA typed samples by NGS; in contrast, the genome wide scan did not include a detailed characterization of the HLA genes. The strongest signals observed in the IHIW samples were in DQA1 and DQB1 in or around exon 2–the portion of the gene responsible for antigen presentation and most likely to be under environmental and evolutionary pressure. Our new statistical approach and analysis suggest novel evolutionary pressure in new regions and additionally highlight the importance of improved sequencing and characterization of variation across the extended MHC and other critical regions.

Introduction section:

Balancing selection occurs when multiple alleles are maintained in a population because of evolutionary factors that affect the fitness of the organism and provide insight into heritable disease and response to pathogenic exposures (Andrés et al. 2009; Bitarello et al. 2018). Neutral alleles that do not affect the survival or reproductivity of an organism are expected to randomly drift with relatively low probability of fixation over many generations. Variants under negative selection will drive allele frequencies down because they are detrimental to the survival of the organism (Davydov et al. 2010; Hayeck et al. 2019; 2020; Gussow et al. 2016), while positive selection (Johnson and Voight 2018; Sabeti et al. 2007; Palamara et al. 2018) may push an allele to fixation if advantageous enough. Balancing selection, on the other hand, may arise from one or multiple sources of evolutionary pressures impacting the allele frequencies including: pleiotropy, overdominance, negative selection, positive selection, or exposure to environmental pressures to name a few; where these competing, or balancing, forces can keep multiple alleles at the same position in equilibrium across a population. Further, as a result, alleles under balancing selection are of interest because they may have one or multiple biological implications, for instance one allele may be related to heritable disease or correspond to specific response to environmental factors, such as a virus. Since multiple alleles are of importance due to the competing evolutionary pressures, multiple alleles are maintained at some equilibrium state. The surrounding variants, on the same haplotypes as the alleles under selection, will also rise to more common frequency (Figure 1), known as hitchhiking, where neighboring variants will then demonstrate stronger than expected association, until disrupted by novel mutations or recombination events (Fay and Wu 2000). This leaves behind the imprint of linkage disequilibrium (LD) blocks, regions that contain strong correlation among neighboring variants, and a higher density of polymorphisms than would be expected from neutral evolutionary genetic drift. So, studying patterns of LD, to find variants more likely to be under selective pressure, helps identify potential regions that impact fitness of humans because they have functional and biologic significance.

The classic example of balancing selection is the environmental pressure imposed by malaria and the heritable impact of sickle cell anemia. Heterozygous carriers at the HBB sickle cell locus are more malaria resistant while homozygous carriers have risk of sickle cell disease and high childhood mortality (Aidoo et al. 2002; Band et al. 2015; Tishkoff et al. 2001; Karlsson, Kwiatkowski, and Sabeti 2014). Early approaches to test for balancing selection alleles focused on detecting enrichment of common alleles, like the HKA statistic (Hudson, Kreitman, and

Aguadé 1987; Wright and Charlesworth 2004), or deviations in the allele site frequency spectrum (SFS) relative to expectation under a neutral evolving population, like Tajima's D (Tajima 1989). Some approaches look to characterize balancing selection that spanned long evolutionary time scales across different species (Siewert and Voight 2020; 2017; Cheng and Degiorgio 2019; DeGiorgio, Lohmueller, and Nielsen 2014; Bitarello et al. 2018); however some of these methods also require the derived versus ancestral allele inferred from ancestral or out of species populations–which may not be available, of good quality, or appropriate for the particular variant. Tests for deviations from neutral drift may miss selective signals. Also, methods that test for trans-species selective alleles may only capture ancient selective signals that affect the fitness of multiple species.

Further, although extensive work has been done looking at the major histocompatibility complex (MHC) as a key example of balancing selection (Lenz et al. 2016; Parham 2005; Barreiro and Quintana-Murci 2010), detailed follow up has been limited because this region is notoriously difficult to sequence and phase accurately. Genes in the MHC are critical to immune function and response, coding for molecules that directly respond to pathogenic exposures, making it a known region of particular interest both from a medical perspective and in terms of better understanding evolutionary pressures. The MHC and corresponding genes are difficult to sequence specifically because of strong LD, high numbers of polymorphisms, and large repetitive regions. Genes found in such highly complex regions are expected to be of particularly interest in trying to characterize patterns of balancing selection. From a clinical standpoint, understanding LD patterns and characterizing variation in polymorphisms in HLA genes are key to understanding immune responses. Detailed sequencing of HLA genes is often needed to provide proper patient-donor matches in transplantation to avoid rejection. In clinical settings, high resolution typing of HLA genes is performed but often is not a part of large consortium analyses.

We developed a new statistical method, a linkage disequilibrium approximate Bayesian factor (LD-ABF) for sparse or fully separated binary data, building on recent evolutionary genetics theory, to investigate patterns of balancing selection by testing for density of polymorphisms and strength of LD then used it to isolate patterns of balancing selection, both genome wide, and a separate follow up analysis across the HLA genes. We curated a novel set of over 500 high coverage clinical samples from trios, phased and processed them, to perform a genome wide scan for selection. This was followed by a focused study on key HLA genes found in a separate dataset, the 17th International HLA and Immunogenetics Workshop (IHIW), a

resource reporting allele frequencies in unrelated world populations based on high resolution sequencing by Next Generation Sequencing (NGS), containing thousands of haplotypes.

Results:

Overview of methods

Variants under balancing selection are expected to leave a pattern of strong LD in close proximity of the mutation and increased levels of polymorphisms. To test for both strong LD and high density of polymorphisms at each SNP, we developed a statistic to determine how well a given variant predicts each of its neighboring polymorphisms. Testing is complicated because we expect to see both rare variants, resulting in sparse data, and settings of close or perfect LD among variants, resulting in quasi or full separation of the data. An effective strategy to address sparse data or full separation is to use penalized regression, which has foundations in both frequentist and Bayesian approaches and in certain settings frequentist penalized estimation will be equivalent to posterior mode estimation (Greenland and Mansournia 2015). To account for these issues, a Bayesian logistic regression model with weakly informative priors, grounded in penalized regression methods, is used since this technique has been shown to be effective in settings of both sparse and fully separated data (Rahman and Sultana 2017; Greenland and Mansournia 2015). Using $logF$ priors, the conjugate family for binomial logistic regression, enables leveraging established data augmentation techniques which is a computationally efficient and robust way to get posterior estimates of the coefficients (Greenland and Mansournia 2015; Greenland 2007; 2003). Then to get a measure of how well the test SNP predicts the neighboring variants, we derived a Bayes factor, which is defined as the ratio of the probability of the null model versus the alternative hypotheses (Kass and Raftery 1995), which can be thought of as somewhat analogous to a frequentist likelihood ratio test and an alternative approach to a frequentist p-value.

The first step is to phase samples in order to make inference on the haplotypes. Depending on the resource, this may either already have been done or can be done using tools like SHAPEIT (Delaneau et al. 2019). The second step is to create augmented logistic regression with pseudo-data and fit a weighted logistic regression model on the pseudo-data, performing this on both the null and alternative model. Third, the estimates of coefficients fit in the augmented logistic regression correspond to the posterior modes for the posterior distributions of the null and alternative models to calculate the approximate Bayes factor (ABF). Fourth, the log of the product of the ABF for every SNP in a window around the test SNP is taken to get the combined LD-ABF, giving both a measure of density of polymorphisms and LD,

and is then divided by the window size to get a scaled statistic. A more detailed derivation of the method can be found in online methods and the code can be found online (https://github.com/tris-10/LD-ABF).

Assessing Prediction of Balancing Selection from Evolutionary Simulations

To test the effectiveness of LD-ABF in detecting balancing selection, a series of forward time simulations were run using using SLiM (Haller and Messer 2017) and the model was fit and compared with other balancing selection test statistics. For each simulation scenario samples of 10,000 were generated across 10 kilobase (Kb) windows, assuming mutation rate and recombination rates of 2.5×10^{-8}. An ancestral population is simulated for 100,000 generations and then a split occurs, then two different balancing selection scenarios are simulated. For older events the balancing selection mutation is introduced at the time of the split in the center of the 10kb region and then another 250,000 generations are simulated. Then for the more recent or younger events the balancing selection mutation is introduced 150,000 generations after the split and then another 100,000 generations are simulated. In either case, the balancing selection variant is introduced at the center of the 10Kb region. Three different equilibrium frequencies were simulated, {0.25,0.5,0.75}, assuming heterozygous fitness of $1 + hs$ using a selection coefficient *s* of 10^{-2} and over dominance coefficient *h* dependent on the desired equilibrium allele frequency corresponding to {-0.5, 100, 1.5}. An equilibrium frequency of 0.75 indicates the derived allele is under enough positive selection that it becomes more common than the ancestral allele. These simulations therefore also indicate some level of detection of positive selection as well–assuming the evolutionary pressure is not a selective sweep that is strong enough to induce full fixation of the allele. For each of the six total scenarios (two-time points vs three equilibrium frequencies) two thousand simulations were run along with an additional neutral set where no balancing selection variant was introduced after the split.

The LD-ABF is compared to the HKA statistic (Hudson, Kreitman, and Aguadé 1987), Tajima's D (Tajima 1989), Beta Scan $\beta_{2,std}$ (Siewert and Voight 2020; 2017), and D_{ng} statistic (Tennessen and Duraisingh 2021). Both HKA and Tajima's D are classic population genetics tests reflecting enrichment in common alleles relative to expectation under a neutral evolving population. Beta Scan looks at a test statistic and compares the weighted regional mutation rate relative to a neutral estimate, while D_{ng} is the sum of the regional binary correlation (Slatkin 2008). All of the statistics appear to perform best for variants of more ancient origin. For example, in the simulations where the balancing selection variant appears 250,000 generations before completion at an equilibrium allele frequency of 50%, the: LD-ABF has an AUC of 98.3%, Tajima's D — 97.7%, D_{ng} — 97.4%, $\beta_{2,std}$ — 96.9%, and HKA — 83.0% (Figure 2 and

Supplemental Figure 1 and Supplemental Table 1). When the balancing selection variant is more recent in origin, the improvement in predictive performance is higher for LD-ABF relative to the other methods: LD-ABF has an AUC of 94.4% Tajima's D— 93.5%, D_{ng} — 92.4%, $\beta_{2,std}$ — 90.5%, and HKA — 68.4%. Although the classic Tajima's D appears second best in several cases, its performance appears inconsistent. When examining more recent variants at an equilibrium frequency of 25%, LD-ABF has an AUC of 93.4%, where Tajima's D has an AUC of 85.8%, corresponding to an AUC improvement of 7.6% with the new method. Generally, most of the methods (besides HKA) appear to perform well, but LD-ABF appears to be the most robust, consistently predicting signals of balancing selection with either the top or within 0.1% of the top AUC in all simulation scenarios.

Genome Wide Scan for Balancing Selection of Clinical Trios

A set of over 500 clinical trios were analyzed, in which SNP array data was augmented with whole exome sequencing (Table 1). The samples were phased using SHAPEIT2 (Choi et al. 2018; Delaneau et al. 2019). Often signals of selection can be obscured or confounded due to demographic shifts across populations, so ancestral populations were inferred to perform within-ancestry analysis. The first 10 principle components (PCs) were calculated from SNP array samples, then K-nearest neighbors algorithm was run using the PCs to infer the closest 1000 Genomes Project (Auton et al. 2015) super population—Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), or the Americas (AMR)—for each individual (Supplemental Figure 2). Then the LD-ABF was calculated across the genome for each of the different populations.

Across each of the populations the top ten peaks were investigated (Figure 3, Table 2 and Supplemental Figure 3 and online data), where LD-ABF scores in the immediate mega-base window around a peak are ignored to determine subsequent peaks, looking at relative gene density and sequence context of each peak. The top peak for the: AFR population is in OR51B5, which is related to sickle cell anemia (Solovieff et al. 2010); while in AMR, EUR, and EAS the top peak is SIRPA, a signal regulatory protein; and for SAS the top peak is in HLA-DPA1, an immune related MHC class II gene. The peaks were initially determined using a conservative window of one mega-base. If that window is reduced to 100 kilobases (online materials and online data) a top 10 LD-ABF score at SIRPB1 is observed in all of the populations (Supplemental Figure 4), where SIRPA and SIRPB1 are part of a "paired receptors" class of membrane proteins (Hatherley et al. 2008). However, this region is known to have homologous sequence and large structural variation, making it difficult to both assemble and align. So, to follow up we performed isolated scans of the SIRP and MHC regions (Figure 4)

using samples from the Pangenome, high quality long read HiFi PacBio data (https://humanpangenome.org/). The Pangenome samples correspond to long read sequencing and are expected to help remove potential artifacts induced by assembly and alignment of other platforms (although a considerably smaller sample size, so not the primary focus of analysis). The signal and long-range LD at SIRPB1 disappears, indicating this is likely an artifact in the CHOP samples. The magnitude of the SIRPA signal though is still greater than the MHC in the Pangenome data, indicating potentially strong selection in that region in a completely separate set of samples. The second strongest selection signal in the AFR samples was observed across SLC35G4, in the SLC35 solute carrier family 35 (SLC35) whose members transport nucleotide sugars (Hadley et al. 2019). As expected, other strongest peaks are in HLA genes. The specific region in each, however, varies from population to population. In the largest populations, EUR and AMR, the top peak is found in HLA-C, while for AFR, EAS, and SAS the top peak is in HLA-DPA1, with additional strong signals of balancing selection seen in MICA in the AFR samples. Over the MHC there appears to be three distinct clusters (Figure 5), which appear to match up with previously noted recombination hotspots found breaking up: (1) HLA-C and HLA-B, (2) HLA-DRB1, HLA-DQA1, and HLA-DQB1, and (3) HLA-DPA1 and HLA-DPB1 (Miretti et al. 2005; Nordin et al. 2020; Cullen et al. 1997). More generally, selection peaks of the LD-ABF appear to occur consistently in immune related genes. In addition to the MHC genes and SIRPA, a peak is found in PSMD13 26S Proteasome, a multicatalytic proteinase found in the EUR samples, where an essential function of the modified proteasome is processing of MHC peptides (Sayers et al. 2019). This gene, PSMD13, passed the same rigorous quality control as the rest of the genome wide scan (See Supplemental Methods) but, should be noted appears close to the telomeric region of chromosome 11 where such regions may have mapping an alignment issues (Supplemental Figure 6). CNR2 is a cannabinoid receptor and has known associations with psychoactive response, but it notably is also related to an anti-inflammatory response (Amberger et al. 2015; 2019). A peak in the Glypican gene, GPC6 (Supplemental Figure 7), was found in all but the EAS populations, while a peak in GBP4, a Guanylate-binding protein coding gene, was found in the EAS samples. CYP412, a peak gene found in several of the populations, encodes a cytochrome protein that catalyze many reactions in drug metabolism and synthesis of cholesterol, steroid, and lipids (Sayers et al. 2019) (Table 2).

Beyond immune related genes, the top LD-ABF peaks lie in or around several olfactory receptor genes, i.e. OR2T4, OR51B5, OR51B6, OR51Q1, OR6K6, OR5P2, and OR51F1, associated with response to smells and typically strong candidates for selection. Similar to the MHC genes with relatively dense gene regions, ranging from 47 to 81 genes within a mega-

base, the olfactory genes with many selection peaks appear to generally be in gene dense regions, ranging from 19 to 49 genes within a mega-base. TAS2R genes, found in top peaks of EUR and SAS, are taste receptors related to sensing bitter tastes and have been thought to be under selective pressure (Kim et al. 2005). Lastly, a couple of Zinc finger protein genes, ZNF717 and ZNF280, are observed in all but the EUR samples.

The top 0.1% of the balancing selection scores were then intersected with known GWAS catalog significant SNPs (Buniello et al. 2019) to find overlap between strong signals of selection and known disease associated variants (Table 3). The strongest LD-ABF score for GWAS significant SNPs in AFR, AMR, and EUR was seen in OR51B5/OR51B6 at rs5006884 with known association with fetal hemoglobin in sickle cell anemia (Solovieff et al. 2010)—this coincides with expectations that the disease is classically found under balancing selection. The strongest signal for EAS was at rs17855611 in SIRPA associated with blood protein levels and for SAS rs1126506 in HLA-DPA1 associated with Anti-rubella virus IgG levels. Many of the traits include blood and immune related diseases, some other traits with observed GWAS significant association were: chronic obstructive pulmonary disease, lumbar spine bone mineral density, folate pathway vitamin levels, pediatric autoimmune diseases, blood protein levels, asthma, Beta-2 microglubulin plasma levels, IgG galactosylation phenotypes, anti-rubella virus IgG levels, antineutrophil cytoplasmic antibody-associated vasculitis, post bronchodilator FEV1, venous thromboembolism, reticulocyte count, late-onset Alzheimer's disease, C-peptide levels in type I diabetes, IgE levels, urinary metabolites (H-NMR features), mean corpuscular hemoglobin concentration, platelet count, intraocular pressure, hemoglobin, red cell distribution width, end-stage coagulation, elevated serum carcinoembryonic antigen levels, obesity-related traits, severe aplastic anemia, and others (Table 3).

It is largely to be expected few or no ClinVar variants (Landrum et al. 2014) would be detected, since the focus of this test is not on rare variation and quality control filter process remove many of them. If we were testing for negative selection, where variants with strong deleterious effects are more expected, then such resource would be expected to have more overlap. However, when decreasing the threshold to include variants in the top 1% several variants related to drug response: several variants within CYP2D6 (rs56011157/rs1081000, rs1080996, rs76312385, rs1080995, rs74644586, rs75276289, rs675026, rs562859) and variants in OPRM1 (rs675026 and rs562859) are related to Tramadol response. The variant kgp11998537 is a risk factor for Systemic lupus erythematosus found in IRF5, and a variant in HAO1 (rs2423326) with association to calcium oxalate urolithiasis (rs2423326).

Receptor genes, cell surface, signaling, and immune genes—especially HLA genes—are all categories of genes expected to be likely candidates for balancing or positive selection because their functionality is directly tied to environmental interactions, such as response to pathogens and sense of smell or taste, and hence they are subject to evolutionary pressures. The HLA genes are known to be difficult to properly sequence because they are highly polymorphic and contain large LD blocks; however, they are of vital importance to the immune system and detailed typing of solely the HLAs is performed in settings like transplantation where matching between patient and donors is critical to avoid organ rejection (Wiebe et al. 2018; Petersdorf et al. 2014; Shieh et al. 2020). Dramatic signatures of selection in chromosome 6, in and around the MHC which contains HLA genes, were observed, including several of the top ten LD-ABF selection peaks across all populations and around 40% of the GWAS significantly associated variants occurring in the MHC. The strong selection signal and the known limitations in sequencing and characterizing the MHC through both SNP array and next generation sequencing technologies motivated further investigation of the region through other sequencing means.

Detailed look at HLA Genes Using High Quality Sequencing

To take a closer look across the HLA genes, haplotypes from the 17th International HLA and Immunogenetics Workshop (IHIW and see online resources), characterized by next generation sequencing in unrelated populations (17 IHIW Table 1), were matched with their established sequences, as described in the interational ImMunoGeneTics (IMGT and see online resources)(Robinson et al. 2020) HLA database version 3.25.0. This consisted of over 3,500 samples (ie over 7,000 HLA alleles per locus), where the exact number depends on the gene, across multiple world populations: European Americans, African Americans, USA Hispanics, Spanish, Mexican, Italian, Greek, Asian Pacific Islanders, Thai, Indian, Arab, and Europeans. Analysis was run on genes containing at least 100 haplotypes for a given population.

The strongest signals are observed in DQA1, DQB1, and DRB1 both in terms of the LD-ABF selection peaks along with the placement of GWAS significant SNPs. The top 10 GWAS associated SNPs were also found across all the different populations as well (and online data). The top LD-ABF SNPs that are GWAS SNPs are associated with traits like: red blood cell count, leukemia, autism, schizophrenia, asthma, and others with sequence context either intronic or missense, which makes sense in the context of looking for signals of balancing selection. The top peaks in LD-ABF signals that were observed in DQB1 (Figure 6 and Supplemental Figure 9) occur in the second exon and appear to extend into the second intron; exon 2 is known to be important in antigen recognition. In DQA1 the highest peak in most populations is in the first

intron (Figure 6), possibly extending out to the second exon, with similar functional implications in the second exon of DQA1. There is a dramatic peak in and around intron 5 for DRB1; however, this peak required further follow up investigation to better understand the potential causes. DRB1 is known to have interesting patterns of extended LD, where depending on the DRB1 haplotype a second expressed DRB gene, either DRB3, DRB4 or DRB5, will typically be observed; however, haplotypes will only contain one of the other DRB genes (Kennedy, Ozbek, and Dorak 2017).

Even though these samples were collected with the specific purpose of characterizing HLA genes it is still expected that some of the non-coding sequence were likely inferred. When primers used to type DRB1 were not capturing intron 5, where the dramatic signal of interest is observed, the sequence was either being inferred or imputed. To further follow up on this, samples from the GeT-RM and Coriell Cell Repositories (Bettinotti et al. 2018) were used to generate complete consensus sequences with proper mapping and alignment over DRB1 was performed to determine if the balancing selection signal was still seen from an external set of samples without using imputation. The dramatic signal observed in the region, where we see a strong LD block within intron 5 (Supplemental Figure 8 and Supplemental Figure 10), intron 5 is known to have an Alu and a LINE–these long transposable elements can make mapping and alignment difficult. The analysis was repeated also looking at the region using Repeat Masker to see if this would impact the inference made across intron 5. When imposing the additional filter, the LD block is essentially removed, causing the balancing selection signal to similarly disappear (Supplemental Figure 8). Other balancing selection studies check filtering on Repeat Masker (Siewert and Voight 2017; Cheng and Degiorgio 2019), with the advantage of removing regions that are both difficult to sequence and possible LD artifacts found from large transposable elements as opposed to regional hitchhiking over long stretches of evolutionary time. While using Repeat Masker to filter has the added benefits as described it comes with the strong caveated disadvantage, Repeat Masker will potentially remove over 20% of variants because these regions are so common. Further follow up was performed to better understand the relationship between LD patterns related to DRB1 in the context of the other DRB genes (online methods).

Discussion section:

Testing for LD in genomic regions is often a reflection of functional interrelationships of the products encoded by these genomic elements and therefore of significance as we explore the role of different genomic regions, whether coding or non-coding. The new method we developed detects signals of evolutionary selective pressures by measuring the strength of LD

and density of variation. The method demonstrates improvement in classification of variants under selection in detailed simulation analysis and builds on theory developed by Siewert and Voight (Siewert and Voight 2017; 2020), while taking a similar approach to Tennessen and Duraisingh (Tennessen and Duraisingh 2021) to specifically look for the density of LD. Previous work has tended to focus only on SNP array samples, utilizing thousand genomes SNP array data. In contrast, here we leverage both SNP array and clinical grade whole exome sequencing data along with improved phasing from trios. Using a curated sample of observational data taken from a clinical diagnostic lab to collect trios, including both SNP array and high coverage whole exome sequencing, along with processing of next generation sequencing typing of thousands of haplotypes from IHIW for further analysis of key HLA genes, helped to better isolate potential signatures of selective pressures.

Based on the genome wide scan across the clinical trios, several novel signals were potentially identified; to our knowledge CYP2A7, GPC6, CNR2, and PSMD13 have not been found to be under selective pressure in previous studies, although related genes or the gene families are consistent with previous findings. SIRPA had one of, if not, the strongest selection signals (depending on the population) which appeared to be consistently on par with the signal in the MHC even in looking in Pangenome long read samples for confirmation (Figure 4). SIRPA is an immune related protein thought to be a key component of the “do-not-eat-me” signaling pathway, where CD47 has been demonstrated to be a ligand for this receptor (Y. P. Chen et al. 2019; Sayers et al. 2019). These SIRP receptor genes are thought to have evolved to counteract pathogen binding (Hatherley et al. 2008). Similar to what was found in the follow up analysis of the HLA genes, the strongest signal appears in or right around the extracellular domains for SIRPA (Bateman et al. 2021). These outward facing residues for SIRPA are analogous to binding of antigens by immunoglobulins and T cells (Barclay 2009). Tennessen et al also observed selection around SIRPA, but they did not appear to identify nearly as strong of a signal as our study. Likely, this in part comes from sample differences and methodological approach but, also a key factor is the platform used to analyze the samples. A limitation of using SNP array data alone is it will be inherently limited in to detect evolutionary selection only on the variants included on the platform, potentially biasing the results.

Consistent patterns of strong LD and density of polymorphisms potentially related to balancing selection or positive selection, based on our test statistic and scan, appeared in or around genetic context related to cell surface function, signaling, receptors, and immune response. This is consistent with previous findings in earlier studies, where evolutionary selection signal has been persistently found across the same gene families and with similar

function as was isolated in this study (Cheng and Degiorgio 2019; Siewert and Voight 2020; DeGiorgio, Lohmueller, and Nielsen 2014; Lenz et al. 2016; Tennessen and Duraisingh 2021). A large proportion of selection signal appears around immune related genes, like the HLA genes and also innate immune regions (Cheng and Degiorgio 2019) like SIRPA (Tennessen and Duraisingh 2021). The other major group of genes, besides the MHC genes, were olfactory receptors (Alonso et al. 2008), genes typically associated with the sense of smell and expected to be under strong selection. Peaks in signals were detected in signal gene groups like Zinc finger proteins (ZNF717 and ZNF280A). Similarly Tennessen and Duraisingh found strong signals around ZNF99, another zinc finger protein (possibly related to viral infection) where the ZNF family is known to be associated with transcriptional regulation, signal transduction, DNA repair, and other processes (Cassandri et al. 2017). Other solute carrier family genes have been found to be potentially under selection (Palamara et al. 2018; Cheng and Degiorgio 2019; Bitarello et al. 2018; Tennessen and Duraisingh 2021), where SLC35G4, identified in this study, has been found to potentially be associated with prostate cancer (W. S. Chen et al. 2020) and is enriched for expression in testes (Uhlén et al. 2015; Thul et al. 2017) (https://www.proteinatlas.org/ENSG00000236396-SLC35G4) implying potential impact on reproductive fitness. A glypican, GPC6, observed with strong selection signal, is a gene which contains SNPs associated with bone mineral density (Kemp et al. 2017) and recessive omodysplasia (Campos-Xavier et al. 2009; Amberger et al. 2019), a rare skeletal dysplasia thought to cause shortening and distal tapering of the humeri and femora (Elçioglu et al. 2004). Having two copies of a given minor allele or compound heterozygous alleles would reduce the fitness of the individual and is a plausible scenario in which to see such selective peaks. Additionally, GPC6 appears to have strong and dense LD over a large block of several mega-bases (Supplemental Figure 7), similar to the MHC. GBP4 had strong selection signal in the EAS samples and has been found to be associated with expression in endothelial cells (Amberger et al. 2019; Tripal et al. 2007). A strong signal seen on chromosome 11 in EUR was around PSMD13, a gene related to antigen processing and presentation (Eleftherohorinou et al. 2011; Sayers et al. 2019), making it a likely candidate for being under selective pressure; it is possibly an entirely novel finding. The MHC has classically been a critical region found to have strong evolutionary selection signal and is also closely tied to different phenotypes. Looking at the intersection of that set with variants already known to have genome wide significantly association (Buniello et al. 2019) with diseases and other phenotypes then restricting to the strongest signals of selection (in the top 0.1% LD-ABF test statistics, Table 3) close to 40% of top came from the MHC. Understanding the evolutionary pressures across the MHC and

accurately characterizing the variation across this region is critical in characterize phenotypes and pathogen response.

Despite combining SNP array data with clinical grade high coverage whole exome sequencing and having the advantage of improved phasing potential by using mostly trios, sequencing quality of regions like the MHC is still expected to be limited. A good source of thousands of high-quality sequenced HLA genes from different world populations data is the 17th IHIW. Utilizing this resource improved the granularity across key HLA genes and interestingly both DQA1 and DQB1 have the strongest signals (outside the possible artifactual signal in DRB1). In contrast, while looking at the genome wide scans across SNP array and exomes of the clinical trios, HLA-C and HLA-DPA1 were the top hits across the MHC in the two largest CHOP populations, EUR and AMR. We believe this highlights the potential limitations of standard sequencing techniques and the need for proper characterization of the MHC. Finding the strongest signal in DQA1 and DQB1 in the IHIW samples is of note because class II molecules present exogenous antigens, meaning they directly interact with foreign bodies such as bacteria or viruses, so seeing stronger signals of balancing selection in class II—in the context of the high-quality sequencing across considerably more samples—appears consistent with expectations that selective pressures are often influenced by environmental factors like pathogens. Looking at DQA1 and DQB1 in more detail, the strong signal appears to be either in or around exon 2 which is known to code for the antigen recognition domain and which could reasonably be explained since polymorphisms in these regions would be directly tied to different pathogenic exposures and corresponding immune response. Heterozygous alleles at these antigen recognition domains increases the repertoire of available pathogens that can be detected in comparison with either of the homozygous allele states.

Several limitations of this work leave room for future investigation. Looking at the genome wide scan of clinical trios, several cytochrome P450 proteins genes were picked up (CYP2A7 a top peak genome wide, CYP2E1 with GWAS significant SNPs, CYP2D6 drug response related ClinVar variants) and these genes are known to be related to metabolism which appears rationale in the context of evolutionary selection. Similar genes in this family have been picked up in other selection studies (Tennessen and Duraisingh 2021); however, these genes are also known to have homologous sequence. For example, CYP2A6 is known to result from crossover with CYP2A7 during recombination (Tanner and Tyndale 2017). Homologous regions pose particular quality control challenges, this was also observed with the strong, and likely artifactual, signal in SIRPB1 (Supplemental Figure 4) that was not replicated using long read data (Figure 4). Although, we did not see excess heterozygotes in this CYP

regions, otherwise a clear indication of mapping issues that was used to filter out such artifacts, we would still caution that signals in genes like these require more extensive study. Future work would be improved by even higher quality sequencing and establishment of multiple reference sequences to distinguish strong LD and density of polymorphisms from structural events, like duplications. Similarly, when just focusing on HLA genes from the IHIW samples, strong signals were observed across DRB1. The strong signal in DRB1 is however likely related to a combination of sequencing and alignment issues in a region known to have strong and long-range LD. This ambiguity suggests the need for better sequencing across the MHC, which, despite being of known critical importance to human health, is often explicitly avoided from interpretation because of its complexity. Other regions of similar complexity to the MHC may also pose challenges in sequencing and interpretation.

The current method focuses on LD within a given window and does not test for long range LD looking at extended haplotypes. In future work we intend to explore possible signal of balancing selection looking at extended haplotypes to better understand the potential of long-range LD, along with gene level patterns of LD as is seen by looking at the DRB genes. The samples from the observational study of clinical trios from CHOP correspond to individuals that have come into the clinical lab likely for some assessment related to disease outcomes. As a result, they do not directly correspond to standing genetic variation in a healthy control sample. However, with the filtering and quality control on the samples, including removal of rare variants, indels, and structural variation which are commonly the focus in diagnostic settings, we believe focusing on the patterns of linkage disequilibrium and common variation in this study is still sound. The ancestry inference for the CHOP samples was performed using standard techniques of mapping individuals to ancestral populations based on relative PC space distance, using the population definitions found in the 1000 Genomes Project. Such inference is expected to have limitations since the samples were not collected prospectively specifically for such a study. As future studies and consortia collect more samples within different populations to directly study genetic variation, analysis methods like LD-ABF will also be expected to improve and grow.

Although there was improvement in signal detection by augmenting SNP array samples with clinical level whole exome sequencing, high coverage whole genome sequencing will likely be part of the next steps to improve interpretation of genetic variation. Short read technologies are inherently limited; any alleles that differ only in sequence further than the maximum read length will result in potential ambiguity in sequencing, which is expected to be relatively common in the MHC in particular. There have been recent advances in sequencing the MHC., in the

future, long read sequencing will become increasingly important (Logsdon, Vollger, and Eichler 2020) to better decipher the complicated patterns and implications of genetic variation in the MHC and other critical regions of the genome. The need for improved sequencing was on display when trying to understand the implications of the strong peak in DRB1 intron 5 and corresponding concordance with other DRB genes. When attempting to identify association between particular polymorphisms and a specific phenotype, because of the LD impact and incomplete sequencing of the MHC, it is difficult to discern the true elements of association. To reconcile these ambiguities, both improved sequencing along with mapping and alignment techniques, like population reference graphs (Dilthey et al. 2016), will be required to create reference genomes for different human populations (Eichler 2019) that will help investigators to discern demographic and ancestral imprints in the genome versus variation implicated in disease response or human traits.

Acknowledgements

We would like to thank both the participants in IHIW17, including the patients who volunteered to have their samples collected and analyzed for research purposes, and the teams across the world that performed the typing and collection of samples. We would also like to thank Steven Pastor for his helpful feedback about the patterns observed over DRB1.

Online Resources
IHIW http://17ihiw.org/17th-ihiw-ngs-hla-data/
IMGT https://www.ebi.ac.uk/ipd/imgt/hla/
NCBI gene database https://www.ncbi.nlm.nih.gov/gene/
Protein Gene Atlas https://www.proteinatlas.org
Pangenome https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC/HG01361/assemblies/
Git repository with code and online data: https://github.com/tris-10/LD-ABF
In addition to the code, data files can be downloaded online data (github Readme.md section Download LD-ABF supplemental files):

- CHOP Trios: Genome Wide LD-ABF test statistics and Peaks detailed for all included populations
- All IHIW: HLA LD-ABF test statistics for all included populations, tab delimited sequence data generated from IHIW and IMGT 3.25, plots across all genes for all included populations

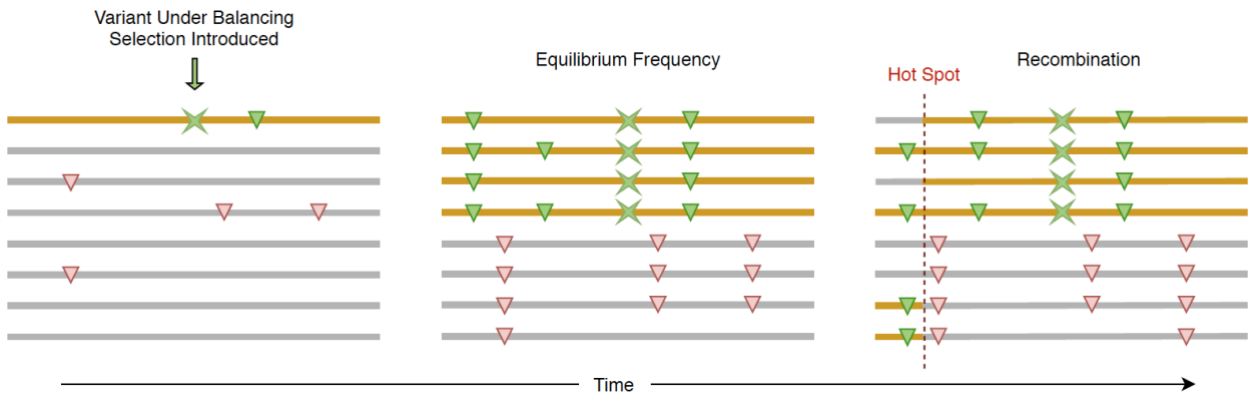

Figure 1 Diagram depicting the progression of an allele under balancing selection, the green X. In the first pane the variant is introduced on a single haplotype. Then after some time has passed evolutionary pressures favoring heterozygotes at the position of focus maintaining both haplotypes with and without the polymorphism, where hitchhiking effects are observed between the variant under balancing selection–inducing LD patterns. Recombination breaks the strong LD resulting in mosaics of the haplotypes, where strong hotspots will diffuse the effects.

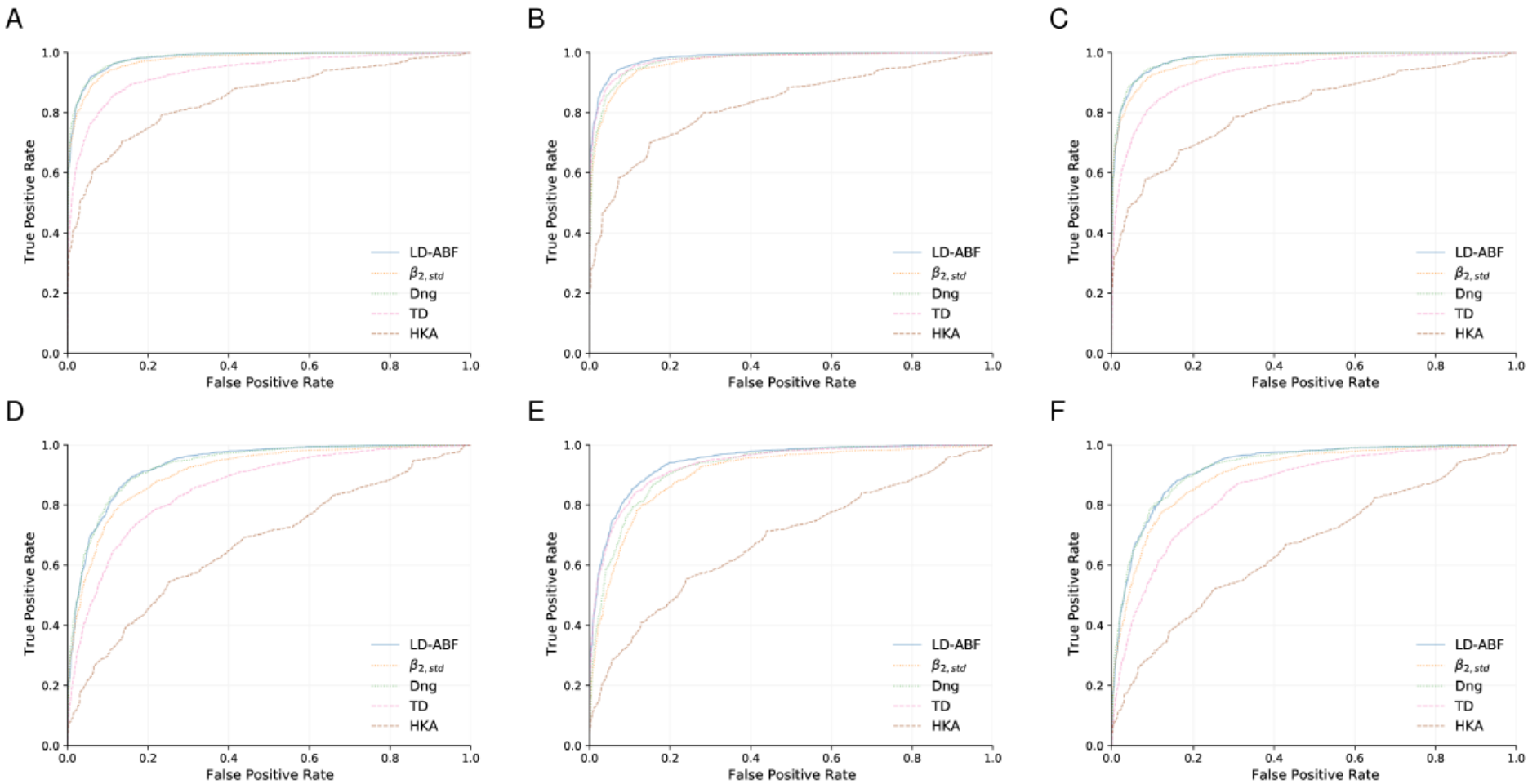

Figure 2 Simulation comparison of methods: ability to detect balancing alleles relative to neutral alleles in different scenarios of equilibrium allele frequency and time when mutation is introduced. *The top row (A, B, C) consists of older mutations (introduced 250,000 generations before completion) versus the bottom row (D, E, F) corresponding to younger mutations (100,000 generations prior to completion) . The left column corresponds to a derived allele frequency of 0.25 (A and D), then 0.5 for the middle (B and E), and 0.75 (C and F) for the right column. True positives are taken from the 2,000 balancing selection simulations for each plot and a random common neutral mutation (with MAF>5%) is used from each of the 2,000 neutral simulations to compare as a false positive, ie each plot corresponds to 2,000 simulations, half with selection and half without, for a given allele frequency and corresponding balancing selection coefficient.*

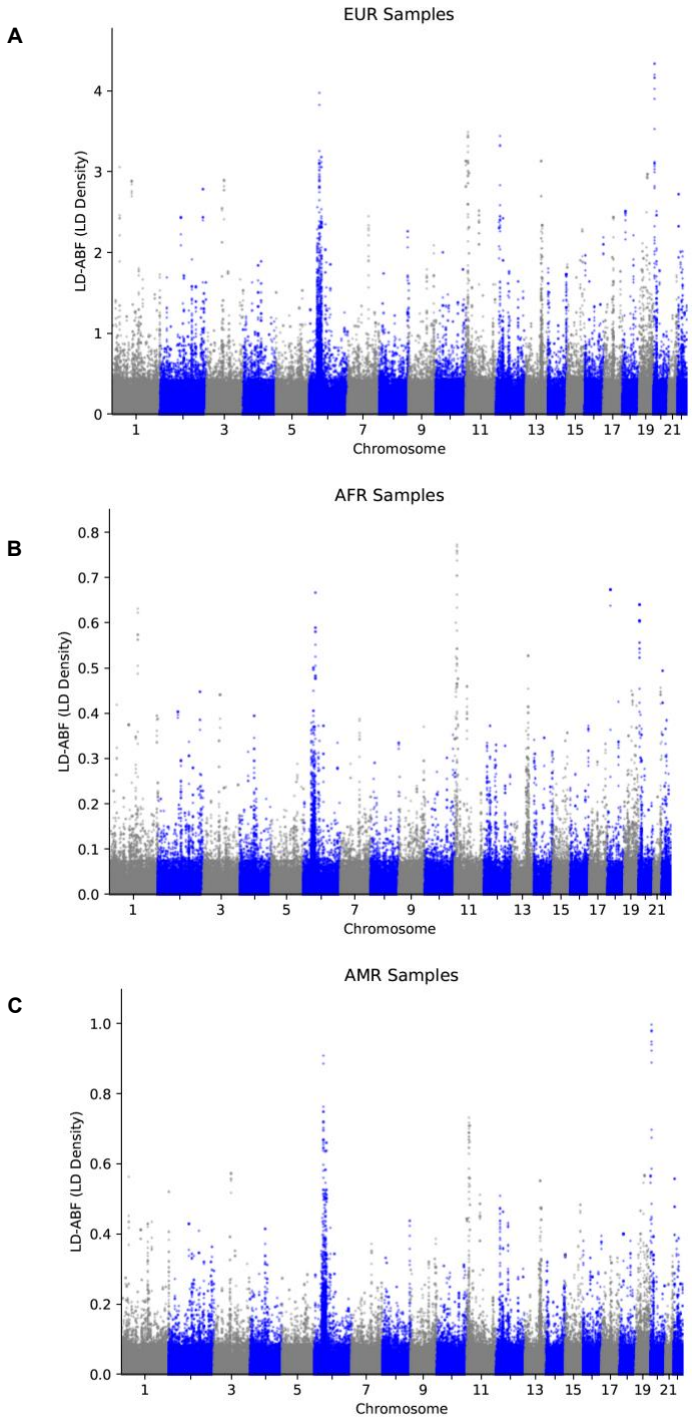

Figure 3 Genome wide scan for balancing selection using LD-ABF, examining 3 different CHOP populations: A.) EUR, B.) AFR, C.) AMR.

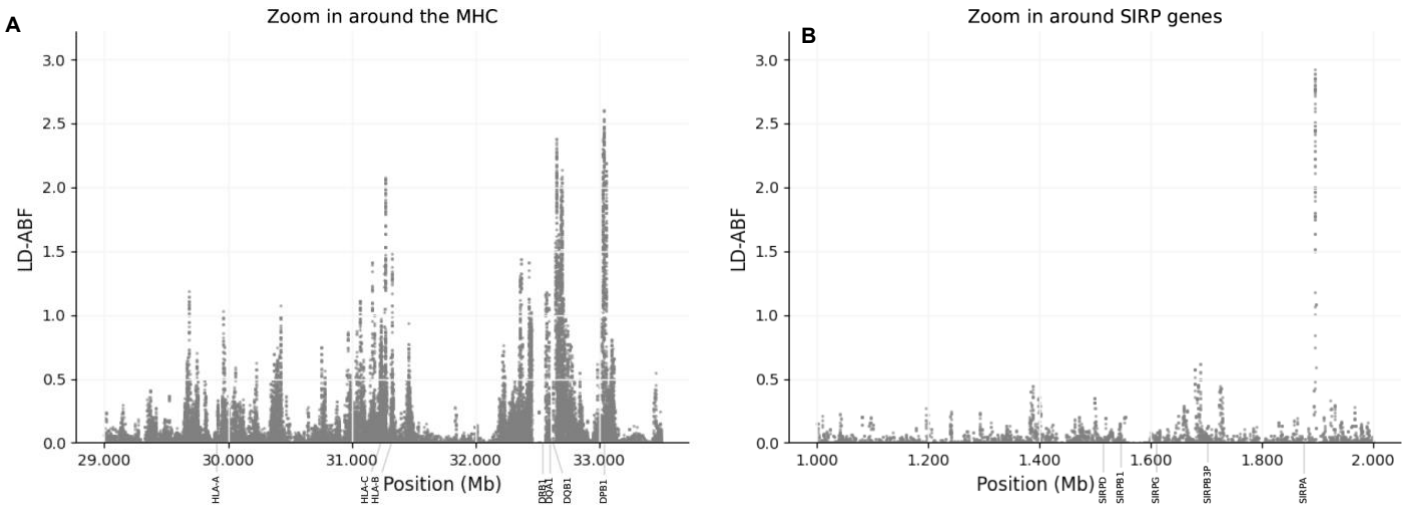

Figure 4 ***Selection signal around the A) MHC and B) SIRP genes looking at Pangenomes PacBio HiFi samples.*** *The start of several genes of interest are labeled in the plots.*

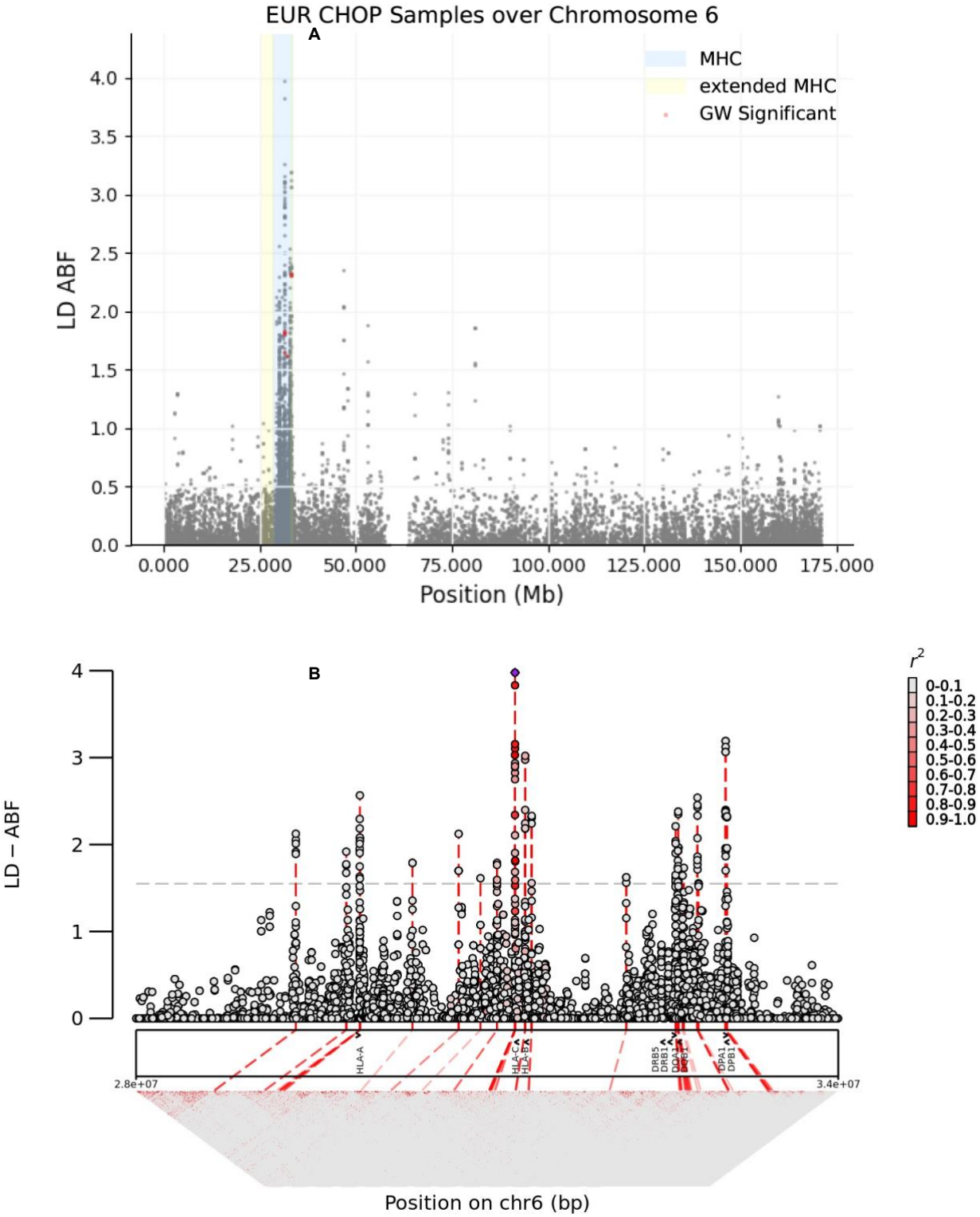

Figure 5 Chromosome 6 with the balancing selection score for the EUR CHOP samples on the y-axis and the position in megabases on the x-axis. The MHC is defined as chr6:28477796-33351542 and the extended MHC is chr6:25726291-33368421. The bottom plot is a zoom on the MHC with class I and II genes labeled and heat map of LD below.

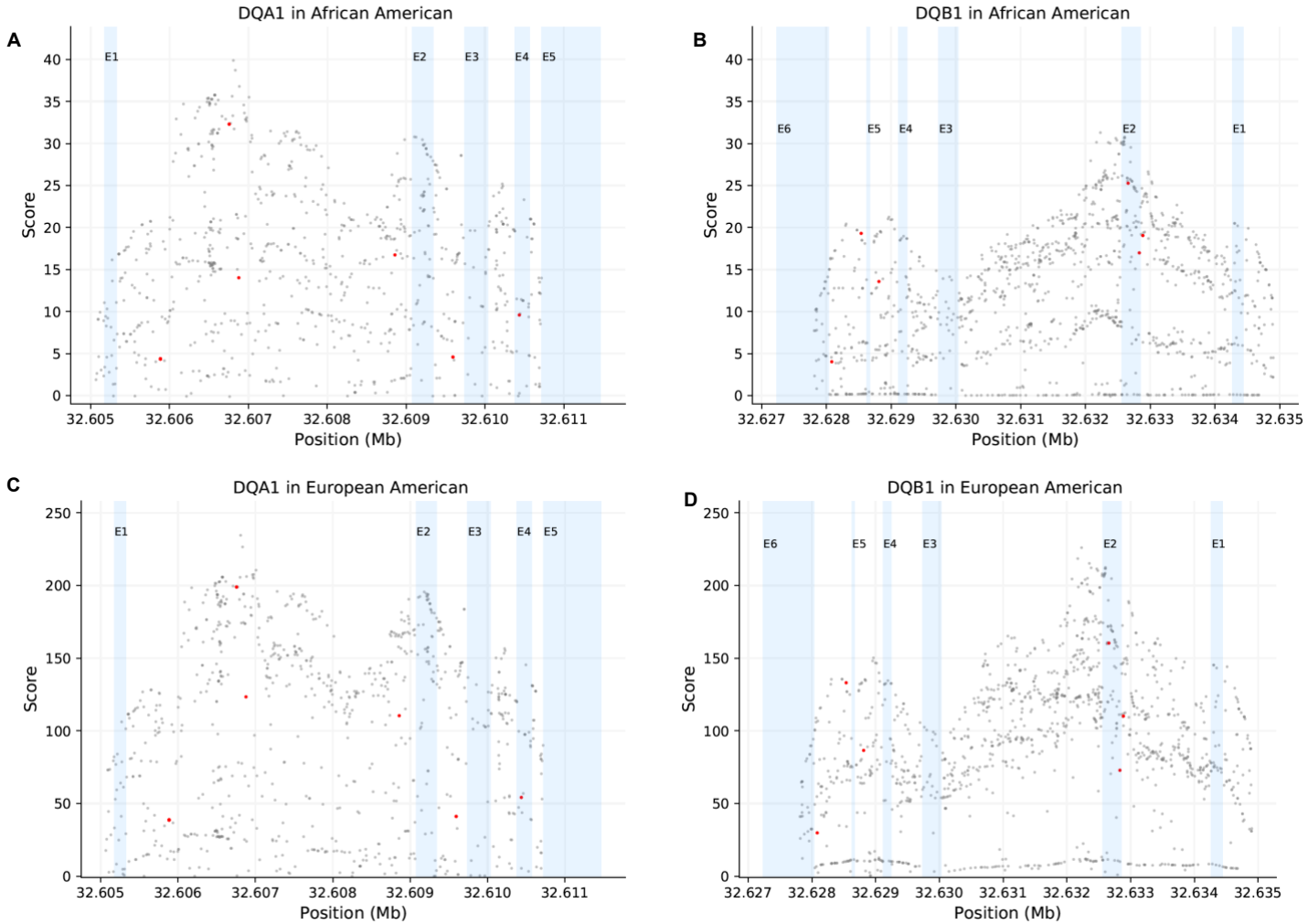

Figure 6 Balancing selection ABF tests for DQA1 and DQB1 from IHIW African American and European American samples. *Looking in the two largest populations in IHIW at the gene HLA-DQA1 and DQB1 we observe a strong signal in each, and around each peak there are known Genome Wide Significant SNPs (red dots) that are associated with blood related traits.*

Population	Individuals	Singleton Probands	Duo Probands	Trio Probands	Probands	Totals
AFR	5	11	10	34	55	60
AMR	5	7	12	44	63	68
EAS	3	8	1	17	26	29
EUR	28	36	34	225	295	323
SAS	2	1	2	18	21	23
Totals	43	63	59	338	460	503

Table 1 Details of population counts for CHOP trios and individuals collected for analysis that include both SNP array data and whole exome sequence data.

Pop	CHROM	ID	LD-ABF	Gene	**Gene Category**	start	end	# Genes Within 100Kb	# Genes Within 1Mb
AFR	11	rs5006884	0.77	OR51B5,OR51B6	Olfactory Receptor Family	5373251	5373251	4	49
	18	rs73943948	0.68	SLC35G4	Solute carrier family	11610121	11610121	2	10
	6	rs34140193	0.67	HLA-DPA1	Major Histocompatibility Complex, Class II	33037080	33037082	3	47
	20	rs6136375	0.64	SIRPA	Signal regulatory protein	1895889	1896100	2	14
	1	rs16841009	0.63	OR6K6	Olfactory Receptor Family	158725194	158725194	4	26
	11	rs11033795	0.60	OR51F1	Olfactory Receptor Family	4790671	4790671	2	38
	13	rs7321055	0.53	GPC6	Glypican	93969248	93969473	0	3
	6	rs1131898	0.50	MICA	Major Histocompatibility Complex, Class I	31379773	31379795	3	81
	22	rs362003	0.50	ZNF280A	Zinc Finger Protein	22869123	22869218	4	12
	11	rs73406603	0.48	OR5P2	Olfactory Receptor Family	7817852	7817959	2	19
AMR	20	rs17855613	1.00	SIRPA	Signal regulatory protein	1895963	1895963	2	14
	6	rs1050147	0.91	HLA-C	Major Histocompatibility Complex, Class I	31237802	31237802	2	73
	11	rs10838093	0.73	OR51B5,OR51Q1	Olfactory Receptor Family	5443887	5443887	5	48
	11	rs78460198	0.71	OR5P2	Olfactory Receptor Family	7817852	7817856	2	19
	6	rs114707319	0.66	HLA-DPA1	Major Histocompatibility Complex, Class II	33037412	33037424	3	47
	3	rs78297221	0.57	ZNF717	Zinc Finger Protein	75786737	75786737	2	8
	19	rs56081734	0.57	CYP2A7	cytochrome P450 proteins	41386420	41386420	4	33
	1	rs3003336	0.57	CNR2	Cannabinoid receptor	24201448	24201448	3	24
	22	rs362132	0.56	ZNF280A	Zinc Finger Protein	22869123	22869218	4	12
	13	rs7321055	0.55	GPC6	Glypican	93969248	93969473	0	3
EUR	20	rs17855616	4.34	SIRPA	Signal regulatory protein	1895963	1895990	2	14
	6	rs9264622	3.98	HLA-C	Major Histocompatibility Complex, Class I	31237876	31237876	2	73
	11	rs5006885	3.50	OR51B5,OR51B6	Olfactory Receptor Family	5373242	5373242	4	49
	12	rs73064966	3.45	PRH1,PRH1-PRR4,PRH1-TAS2R14,TAS2R43	heterogeneous family of proline-rich salivary glycoproteins, taste receptor	11244390	11244390	3	31
	6	rs1126543	3.19	HLA-DPA1	Major Histocompatibility Complex, Class II	33037419	33037424	3	47
	13	rs7338482	3.14	GPC6	Glypican	93969248	93969473	0	3
	11	rs10902112	3.14	PSMD13	26S Proteasome, a multicatalytic proteinase	244106	244167	9	36
	1	rs3003336	3.06	CNR2	Cannabinoid receptor	24201448	24201448	3	24
	11	rs73406603	2.99	OR5P2	Olfactory Receptor Family	7817852	7817959	2	19
	19	rs56081734	2.98	CYP2A7	cytochrome P450 proteins	41386420	41386420	4	33
EAS	20	rs17855615	0.42	SIRPA	Signal regulatory protein	1895889	1896060	2	14
	6	rs114707319	0.37	HLA-DPA1	Major Histocompatibility Complex, Class II	33037412	33037412	3	47
	6	rs1050147	0.30	HLA-C	Major Histocompatibility Complex, Class I	31237802	31237802	2	73
	11	rs10838093	0.27	OR51B5,OR51Q1	Olfactory Receptor Family	5443887	5443887	5	48
	1	rs61248663	0.26	OR2T4	Olfactory Receptor Family	248525328	248525330	5	37
	1	rs1142889	0.24	GBP4	Guanylate-binding proteins	89652071	89652090	2	16
	22	rs362124	0.24	ZNF280A	Zinc Finger Protein	22869123	22869218	4	12
	14	rs200159018	0.23	AHNAK2	encodes a large nucleoprotein	105418234	105418235	3	27
	1	rs3003336	0.23	CNR2	Cannabinoid receptor	24201448	24201448	3	24
	6	rs402388	0.22	FNDC1	fibronectin type III domain containing 1	159654994	159654994	1	13
SAS	6	rs1126542	0.33	HLA-DPA1	Major Histocompatibility Complex, Class II	33037412	33037424	3	47
	20	rs72620874;rs1057114	0.30	SIRPA	Signal regulatory protein	1895889	1895889	2	14
	12	rs73064966	0.24	PRH1,PRH1-PRR4,PRH1-TAS2R14,TAS2R43	heterogeneous family of proline-rich salivary glycoproteins, taste receptor	11244378	11244390	3	31
	1	rs56721857	0.21	OR2T4	Olfactory Receptor Family	248525328	248525330	5	37
	11	rs10838093	0.20	OR51B5,OR51Q1	Olfactory Receptor Family	5443887	5443887	5	48
	1	rs3003336	0.19	CNR2	Cannabinoid receptor	24201448	24201448	3	24
	19	rs4142867	0.18	CYP2A7	cytochrome P450 proteins	41386136	41386136	4	33
	2	rs3755320	0.18	UGT1A10,UGT1A5,UGT1A6,UGT1A7,UGT1A8,UGT1A9	UDP-glucuronosyltransferase	234622061	234622110	9	22
	22	rs362003	0.17	ZNF280A	Zinc Finger Protein	22869123	22869218	4	12
	13	rs7339375	0.17	GPC6	Glypican	93969248	93969473	0	3

Table 2 Top genome wide peaks in balancing selection signal in CHOP trios genic context and regional density. *Often the exact maximum would occur at multiple variants in perfect LD around a particular region. The start and end positions are represented here and the details for space while each individual variant can be found in the online materials.*

Pop	CHROM	ID	LD-ABF	Genes	DISEASE/TRAIT	CONTEXT
AFR	11	rs5006884	0.77	OR51B5, OR51B6	Fetal hemoglobin levels in sickle cell anemia	Missense
	6	rs9277354	0.48	HLA-DPB1	Antineutrophil cytoplasmic antibody-associated vasculitis	frameshift
	6	rs9277356	0.48	HLA-DPB1	Response to hepatitis B vaccine	Missense
	6	rs1126506	0.48	HLA-DPA1	Anti-rubella virus IgG levels	Splice region
	20	rs17855611	0.45	SIRPA	Blood protein levels	Missense
	2	rs4988958	0.40	IL1RI1, IL1RL2, IL18R1	Asthma (childhood onset)	Synonymous
	19	rs35534776	0.37	LILRA1	Blood protein levels	Missense
	6	rs520692	0.31	C4A	Feeling worry	Missense
	6	rs611779	0.31	ADGRF5	Blood protein levels	Synonymous
	6	rs1042151	0.30	HLA-DPB1	Aspirin exacerbated respiratory disease in asthmatics, Severe aplastic anemia	Missense
	19	rs602662	0.29	FUT2	Folate pathway vitamin levels, Pediatric autoimmune diseases	Missense
	10	rs2249694	0.27	CYP2E1	Obesity-related traits	Intron
	6	rs2858331	0.27	HLA-DQA2	IgE levels	Regulatory region
	6	rs422951	0.25	NOTCH4	Chronic hepatitis B infection	Missense
AMR	11	rs5006884	0.71	OR51B5, OR51B6	Fetal hemoglobin levels in sickle cell anemia	Missense
	6	rs2894204	0.53	HLA-C	Waist-hip ratio	Intron
	6	rs1126506	0.53	HLA-DPA1	Anti-rubella virus IgG levels	Splice region
	6	rs9277354	0.51	HLA-DPB1	Antineutrophil cytoplasmic antibody-associated vasculitis	Frameshift
	6	rs9277356	0.51	HLA-DPB1	Response to hepatitis B vaccine	Missense
	20	rs17855611	0.44	SIRPA	Blood protein levels	Missense
	6	rs9264638	0.43	HLA-C	Beta-2 microglubulin plasma levels	Intron
	2	rs4988958	0.43	IL1RI1, IL1RL2, IL18R1	Asthma (childhood onset)	Synonymous
	1	rs4525	0.39	F5	Blood protein levels	Missense
	6	rs1050451	0.38	HLA-B, HLA-C	IgG galactosylation phenotypes (multivariate analysis)	Missense
	6	rs34794906	0.37	HLA-C	Reticulocyte count	Synonymous
	1	rs4524	0.36	F5	Venous thromboembolism	Missense
	6	rs2516703	0.33	HCG17	Itch intensity from mosquito bite	Intron
	19	rs602662	0.29	FUT2	Folate pathway vitamin levels, Pediatric autoimmune diseases	Missense
	20	rs3790160	0.32	JAG1	Lumbar spine bone mineral density	Intron
	19	rs35534776	0.31	LILRA1	Blood protein levels	Missense
EUR	11	rs5006884	3.44	OR51B5, OR51B6	Fetal hemoglobin levels in sickle cell anemia	Missense
	2	rs4988958	2.44	IL1RI1, IL1RL2, IL18R1	Asthma (childhood onset)	Synonymous
	6	rs9277354	2.32	HLA-DPB1	Antineutrophil cytoplasmic antibody-associated vasculitis	Frameshift
	6	rs9277356	2.32	HLA-DPB1	Response to hepatitis B vaccine	Missense
	6	rs1126506	2.31	HLA-DPA1	Anti-rubella virus IgG levels	Splice region
	6	rs2894204	1.83	HLA-C	Waist-hip ratio	Intron
	6	rs1050451	1.81	HLA-B, HLA-C	IgG galactosylation phenotypes (multivariate analysis)	Missense
	17	rs1864325	1.80	MAPT	Lumbar spine bone mineral density	Intron
	17	rs12373142	1.78	SPPL2C	Chronic obstructive pulmonary disease	Missense
	20	rs17855611	1.70	SIRPA	Blood protein levels	Missense
	19	rs602662	0.29	FUT2	Folate pathway vitamin levels, Pediatric autoimmune diseases	Missense
	6	rs9264638	1.66	HLA-C	Beta-2 microglubulin plasma levels	Intron
	8	rs56117011	1.63	PLEC	Post bronchodilator FEV1	Synonymous
	6	rs520692	1.62	C4A	Feeling worry	Missense
	8	rs35916068	1.62	PLEC	Post bronchodilator FEV1	Synonymous
	1	rs4525	1.56	F5	Blood protein levels	Missense

Table 3 top Balancing selection GWAS significant SNPs with strong signals of selection in the top 0.1% looking at the CHOP trios for the EUR, AFR, and AMR populations (continued in Supplemental Table 2).

Population	Category	CHROM	ID	LD ABF	GENE	CLNDN	MC
AFR	Risk Factor	7	kgp11998537	0.11	IRF5	Systemic lupus erythematosus	3 prime UTR
AMR	Drug Response	22	rs56011157;rs1081000	0.15	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs75276289	0.15	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs76312385	0.15	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs74644586	0.15	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs1080996	0.15	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs1080995	0.15	CYP2D6	Tramadol response	Intronic
	Risk Factor	7	kgp11998537	0.13	IRF5	Systemic lupus erythematosus	3 prime UTR
EAS	Risk Factor	7	kgp11998537	0.05	IRF5	Systemic lupus erythematosus	3 prime UTR
EUR	Drug Response	22	rs56011157;rs1081000	0.79	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs1080996	0.78	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs76312385	0.78	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs1080995	0.78	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs74644586	0.76	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs75276289	0.72	CYP2D6	Tramadol response	Intronic
	Drug Response	6	rs675026	0.69	OPRM1	Tramadol response	Synonymous, Intronic
	Drug Response	6	rs562859	0.69	OPRM1	Tramadol response	Synonymous, Intronic
SAS	Association	20	rs2423326	0.05	HAO1	Calcium oxalate urolithiasis	Intronic
	Drug Response	22	rs56011157;rs1081000	0.04	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs75276289	0.04	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs76312385	0.04	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs74644586	0.04	CYP2D6	Tramadol response	Intronic
	Drug Response	22	rs1080996	0.04	CYP2D6	Tramadol response	Intronic

Table 4 variants in the top 1% of LD ABF looking at CHOP clinical Trios with clinical interpretations from ClinVar.

Population	Gene	POS	ID	ABF	DISEASE/TRAIT	CONTEXT	PUBMEDID
European American	DQA1	32606756	rs9272535	198.97	Red blood cell count	Missense	27863252
European American	DQA1	32606756	rs9272535	198.97	Chronic lymphocytic leukemia	Missense	21131588
European American	DQB1	32632659	rs9274390	160.36	Autism spectrum disorder or schizophrenia	Missense	28540026
European American	DRB1	32556601	rs28724212	159.45	Autism spectrum disorder or schizophrenia	Intronic	28540026
European American	DQB1	32628538	rs201043192	133.09	Lung function (low FEV1 vs high FEV1)	Intronic	26423011
European American	DRB1	32550322	rs9269853	128.65	Alzheimer's disease (late onset)	Intronic	30617256
European American	DRB1	32552095	rs17885382	124.85	Asparaginase hypersensitivity in acute lymphoblastic leukemia	Missense	25987655
European American	DQA1	32606878	rs9272544	123.38	Granulocyte percentage of myeloid white cells	Missense	27863252
European American	DRB1	32554129	rs9270074	114.73	Autism spectrum disorder or schizophrenia	Intronic	28540026
European American	DQA1	32608858	rs4455710	110.42	Squamous cell carcinoma	Intronic	26829030
European American	DQB1	32632887	rs201184533	110.13	Asthma	Intronic	30929738
African American	DQA1	32606756	rs9272535	32.30	Red blood cell count	Missense	27863252
African American	DQA1	32606756	rs9272535	32.30	Chronic lymphocytic leukemia	Missense	21131588
African American	DQB1	32632659	rs9274390	25.27	Autism spectrum disorder or schizophrenia	Missense	28540026
African American	DQB1	32628538	rs201043192	19.30	Lung function (low FEV1 vs high FEV1)	Intronic	26423011
African American	DQB1	32632887	rs201184533	19.04	Asthma	Intronic	30929738
African American	DQB1	32632832	rs9274407	16.98	Drug-induced liver injury (amoxicillin-clavulanate)	Missense	21570397
African American	DRB1	32550322	rs9269853	16.96	Alzheimer's disease (late onset)	Intronic	30617256
African American	DQA1	32608858	rs4455710	16.74	Squamous cell carcinoma	Intronic	26829030
African American	DRB1	32554129	rs9270074	16.72	Autism spectrum disorder or schizophrenia	Intronic	28540026
African American	DRB1	32556601	rs28724212	14.76	Autism spectrum disorder or schizophrenia	Intronic	28540026
African American	DQA1	32606878	rs9272544	14.02	Granulocyte percentage of myeloid white cells	Missense	27863252

Table 5 the top GWAS significant SNPs with strong signals of balancing selection found in HLA genes within European American and African American populations from IHIW samples.

Bibliography

Aidoo, Michael, Dianne J. Terlouw, Margarette S. Kolczak, Peter D. McElroy, Feiko O. Ter Kuile, Simon Kariuki, Bernard L. Nahlen, Altaf A. Lal, and Venkatachalam Udhayakumar. 2002. "Protective Effects of the Sickle Cell Gene against Malaria Morbidity and Mortality." *Lancet* 359 (9314): 1311–12. https://doi.org/10.1016/S0140-6736(02)08273-9.

Alonso, Santos, Saioa López, Neskuts Izagirre, and Concepción De La Rúa. 2008. "Overdominance in the Human Genome and Olfactory Receptor Activity." *Molecular Biology and Evolution* 25 (5): 997–1001. https://doi.org/10.1093/molbev/msn049.

Amberger, Joanna S., Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. 2015. "OMIM.Org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders." *Nucleic Acids Research* 43 (D1): D789–98. https://doi.org/10.1093/nar/gku1205.

Amberger, Joanna S., Carol A. Bocchini, Alan F. Scott, and Ada Hamosh. 2019. "OMIM.Org: Leveraging Knowledge across Phenotype-Gene Relationships." *Nucleic Acids Research* 47 (D1): D1038–43. https://doi.org/10.1093/nar/gky1151.

Amemiya, Haley M., Anshul Kundaje, and Alan P. Boyle. 2019. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome." *Scientific Reports* 9 (1): 1–5. https://doi.org/10.1038/s41598-019-45839-z.

Andrés, Aida M., Melissa J. Hubisz, Amit Indap, Dara G. Torgerson, Jeremiah D. Degenhardt, Adam R. Boyko, Ryan N. Gutenkunst, et al. 2009. "Targets of Balancing Selection in the Human Genome." *Molecular Biology and Evolution* 26 (12): 2755–64. https://doi.org/10.1093/molbev/msp190.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Band, Gavin, Kirk A. Rockett, Chris C.A. Spencer, Dominic P. Kwiatkowski, Quang Si Le, Geraldine M. Clarke, Katja Kivinen, et al. 2015. "A Novel Locus of Resistance to Severe Malaria in a Region of Ancient Balancing Selection." *Nature* 526 (7572): 253–57. https://doi.org/10.1038/nature15390.

Barclay, A. Neil. 2009. "Signal Regulatory Protein Alpha (SIRPα)/CD47 Interaction and Function." *Current Opinion in Immunology* 21 (1): 47–52. https://doi.org/10.1016/j.coi.2009.01.008.

Barreiro, Luis B., and Lluís Quintana-Murci. 2010. "From Evolutionary Genetics to Human Immunology: How Selection Shapes Host Defence Genes." *Nature Reviews Genetics* 11 (1): 17–30. https://doi.org/10.1038/nrg2698.

Bateman, Alex, Maria Jesus Martin, Sandra Orchard, Michele Magrane, Rahat Agivetova, Shadab Ahmad, Emanuele Alpi, et al. 2021. "UniProt: The Universal Protein Knowledgebase in 2021." *Nucleic Acids Research* 49 (D1): D480–89. https://doi.org/10.1093/nar/gkaa1100.

Bettinotti, Maria P., Deborah Ferriola, Jamie L. Duke, Timothy L. Mosbruger, Nikolaos Tairis, Lawrence Jennings, Lisa V. Kalman, and Dimitri Monos. 2018. "Characterization of 108 Genomic DNA Reference Materials for 11 Human Leukocyte Antigen Loci: A GeT-RM Collaborative Project." *Journal of Molecular Diagnostics* 20 (5): 703–15. https://doi.org/10.1016/j.jmoldx.2018.05.009.

Bitarello, Barbara D., Cesare De Filippo, João C. Teixeira, Joshua M. Schmidt, Philip Kleinert, Diogo Meyer, and Aida M. Andres. 2018. "Signatures of Long-Term Balancing Selection in Human Genomes." *Genome Biology and Evolution* 10 (3): 939–55. https://doi.org/10.1093/gbe/evy054.

Buniello, Annalisa, Jacqueline A.L. Macarthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. "The NHGRI-EBI GWAS Catalog of

Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." *Nucleic Acids Research* 47 (D1): D1005–12. https://doi.org/10.1093/nar/gky1120.
Campos-Xavier, Ana Belinda, Danielle Martinet, John Bateman, Dan Belluoccio, Lynn Rowley, Tiong Yang Tan, Alica Baxová, et al. 2009. "Mutations in the Heparan-Sulfate Proteoglycan Glypican 6 (GPC6) Impair Endochondral Ossification and Cause Recessive Omodysplasia." *American Journal of Human Genetics* 84 (6): 760–70. https://doi.org/10.1016/j.ajhg.2009.05.002.
Canario, Remy. 2020. "The Best Classifier for Small Datasets : Log- F (m , m) Logit." *Medium*, 1–15. https://medium.com/@remycanario17/log-f-m-m-logit-the-best-classification-algorithm-for-small-datasets-fc92fd95bc58.
Cassandri, Matteo, Artem Smirnov, Flavia Novelli, Consuelo Pitolli, Massimiliano Agostini, Michal Malewicz, Gerry Melino, and Giuseppe Raschellà. 2017. "Zinc-Finger Proteins in Health and Disease." *Cell Death Discovery* 3 (1). https://doi.org/10.1038/cddiscovery.2017.71.
Chen, Ming Huei, Laura M. Raffield, Abdou Mousas, Saori Sakaue, Jennifer E. Huffman, Arden Moscati, Bhavi Trivedi, et al. 2020. *Trans-Ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations*. *Cell*. Vol. 182. https://doi.org/10.1016/j.cell.2020.06.045.
Chen, William S., Winston A. Haynes, Rebecca Waitz, Kathy Kamath, Agustin Vega-Crespo, Raunak Shrestha, Minlu Zhang, et al. 2020. "Autoantibody Landscape in Patients with Advanced Prostate Cancer." *Clinical Cancer Research* 26 (23): 6204–14. https://doi.org/10.1158/1078-0432.CCR-20-1966.
Chen, Ya Ping, Hyo Jin Kim, Hongyan Wu, Tammy Price-Troska, Jose C. Villasboas, Shahrzad Jalali, Andrew L. Feldman, Anne J. Novak, Zhi Zhang Yang, and Stephen M. Ansell. 2019. "SIRPα Expression Delineates Subsets of Intratumoral Monocyte/Macrophages with Different Functional and Prognostic Impact in Follicular Lymphoma." *Blood Cancer Journal* 9 (10). https://doi.org/10.1038/s41408-019-0246-0.
Cheng, Xiaoheng, and Michael Degiorgio. 2019. "Detection of Shared Balancing Selection in the Absence of Trans-Species Polymorphism." *Molecular Biology and Evolution* 36 (1): 177–99. https://doi.org/10.1093/molbev/msy202.
Choi, Yongwook, Agnes P. Chan, Ewen Kirkness, Amalio Telenti, and Nicholas J. Schork. 2018. "Comparison of Phasing Strategies for Whole Human Genomes." *PLOS Genetics* 14 (4): e1007308. https://doi.org/10.1371/journal.pgen.1007308.
Cullen, Michael, Janelle Noble, Henry Erlich, Karen Thorpe, Stephan Beck, William Klitz, John Trowsdale, and Mary Carrington. 1997. "Characterization of Recombination in the HLA Class 11 Region," 397–407.
Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. "Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++." *PLoS Computational Biology* 6 (12). https://doi.org/10.1371/journal.pcbi.1001025.
DeGiorgio, Michael, Kirk E. Lohmueller, and Rasmus Nielsen. 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.
Delaneau, Olivier, Jean-François Zagury, Matthew R. Robinson, Jonathan L. Marchini, and Emmanouil T. Dermitzakis. 2019. "Accurate, Scalable and Integrative Haplotype Estimation." *Nature Communications 2019 10:1* 10 (1): 1–10. https://doi.org/10.1038/s41467-019-13225-y.
Dilthey, Alexander T., Pierre Antoine Gourraud, Alexander J. Mentzer, Nezih Cereb, Zamin Iqbal, and Gil McVean. 2016. "High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs." *PLoS Computational Biology* 12 (10): 1–17. https://doi.org/10.1371/journal.pcbi.1005151.

Eichler, Evan E. 2019. “Genetic Variation, Comparative Genomics, and the Diagnosis of Disease.” *New England Journal of Medicine* 381 (1): 64–74. https://doi.org/10.1056/NEJMra1809315.

Elçioglu, Nursel H., Karl H. Gustavson, Andrew O.M. Wilkie, Memune Yüksel-Apak, and Jürgen W. Spranger. 2004. “Recessive Omodysplasia: Five New Cases and Review of the Literature.” *Pediatric Radiology* 34 (1): 75–82. https://doi.org/10.1007/s00247-003-1064-9.

Eleftherohorinou, Hariklia, Clive J. Hoggart, Victoria J. Wright, Michael Levin, and Lachlan J.M. Coin. 2011. “Pathway-Driven Gene Stability Selection of Two Rheumatoid Arthritis GWAS Identifies and Validates New Susceptibility Genes in Receptor Mediated Signalling Pathways.” *Human Molecular Genetics* 20 (17): 3494–3506. https://doi.org/10.1093/hmg/ddr248.

Fay, Justin C., and Chung I. Wu. 2000. “Hitchhiking under Positive Darwinian Selection.” *Genetics* 155 (3): 1405–13.

Greenland, Sander. 2003. “Generalized Conjugate Priors for Bayesian Analysis of Risk and Survival Regressions.” *Biometrics* 59 (1): 92–99. https://doi.org/10.1111/1541-0420.00011.

———. 2007. “Bayesian Perspectives for Epidemiological Research. II. Regression Analysis.” *International Journal of Epidemiology* 36 (1): 195–202. https://doi.org/10.1093/ije/dyl289.

Greenland, Sander, and Mohammad Ali Mansournia. 2015. “Penalization, Bias Reduction, and Default Priors in Logistic and Related Categorical and Survival Regressions.” *Statistics in Medicine* 34 (23): 3133–43. https://doi.org/10.1002/sim.6537.

Gussow, Ayal B., Slavé Petrovski, Quanli Wang, Andrew S. Allen, and David B. Goldstein. 2016. “The Intolerance to Functional Genetic Variation of Protein Domains Predicts the Localization of Pathogenic Mutations within Genes.” *Genome Biology* 17 (1): 9. https://doi.org/10.1186/s13059-016-0869-4.

Hadley, Barbara, Thomas Litfin, Chris J. Day, Thomas Haselhorst, Yaoqi Zhou, and Joe Tiralongo. 2019. “Nucleotide Sugar Transporter SLC35 Family Structure and Function.” *Computational and Structural Biotechnology Journal* 17: 1123–34. https://doi.org/10.1016/j.csbj.2019.08.002.

Haller, Benjamin C., and Philipp W. Messer. 2017. “SLiM 2: Flexible, Interactive Forward Genetic Simulations.” *Molecular Biology and Evolution* 34 (1): 230–40. https://doi.org/10.1093/molbev/msw211.

Hartl, Daniel L., and Andrew G. Clark. 2007. *Principles Of Population Genetics*. https://doi.org/10.1002/(SICI)1521-1878(199812)20:12<1055::AID-BIES14>3.0.CO;2-X.

Hatherley, Deborah, Stephen C. Graham, Jessie Turner, Karl Harlos, David I. Stuart, and A. Neil Barclay. 2008. “Paired Receptor Specificity Explained by Structures of Signal Regulatory Proteins Alone and Complexed with CD47.” *Molecular Cell* 31 (2): 266–77. https://doi.org/10.1016/j.molcel.2008.05.026.

Hayeck, Tristan J., Nicholas Stong, Charles J. Wolock, Brett Copeland, Sitharthan Kamalakaran, David B. Goldstein, and Andrew S. Allen. 2019. “Improved Pathogenic Variant Localization via a Hierarchical Model of Sub-Regional Intolerance.” *The American Journal of Human Genetics*, 1–11. https://doi.org/10.1016/j.ajhg.2018.12.020.

Hayeck, Tristan J, Nicholas Stong, Evan Baugh, Ryan Dhindsa, Tychele N Turner, Ayan Malakar, and Yuncheng Duan. 2020. “Ancestry Adjustment Improves Genome-Wide Estimates of Regional Intolerance.” https://doi.org/10.1101/2020.03.05.979203.

Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. “A Test of Neutral Molecular Evolution Based on Nucleotide Data.” *Genetics* 116 (1): 153–59.

Johnson, Kelsey Elizabeth, and Benjamin F. Voight. 2018. “Patterns of Shared Signatures of Recent Positive Selection across Human Populations.” *Nature Ecology and Evolution* 2 (4): 713–20. https://doi.org/10.1038/s41559-018-0478-6.

Karlsson, Elinor K., Dominic P. Kwiatkowski, and Pardis C. Sabeti. 2014. “Natural Selection and Infectious Disease in Human Populations.” *Nature Reviews Genetics* 15 (6): 379–93.

https://doi.org/10.1038/nrg3734.
Kass, Robert E., and Adrian E. Raftery. 1995. “Bayes Factors.” *Journal of the American Statistical Association* 90 (430): 773–95. https://doi.org/10.1080/01621459.1995.10476572.
Kemp, John P., John A. Morris, Carolina Medina-Gomez, Vincenzo Forgetta, Nicole M. Warrington, Scott E. Youlten, Jie Zheng, et al. 2017. “Identification of 153 New Loci Associated with Heel Bone Mineral Density and Functional Involvement of GPC6 in Osteoporosis.” *Nature Genetics* 49 (10): 1468–75. https://doi.org/10.1038/ng.3949.
Kennedy, A. E., U. Ozbek, and M. T. Dorak. 2017. “What Has GWAS Done for HLA and Disease Associations?” *International Journal of Immunogenetics* 44 (5): 195–211. https://doi.org/10.1111/iji.12332.
Kim, Unkyung, Stephen Wooding, Dante Ricci, Lynn B. Jorde, and Dennis Drayna. 2005. “Worldwide Haplotype Diversity and Coding Sequence Variation at Human Bitter Taste Receptor Loci.” *Human Mutation* 26 (3): 199–204. https://doi.org/10.1002/humu.20203.
Landrum, Melissa J., Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. 2014. “ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype.” *Nucleic Acids Research* 42 (D1): 980–85. https://doi.org/10.1093/nar/gkt1113.
Lenz, Tobias L., Victor Spirin, Daniel M. Jordan, and Shamil R. Sunyaev. 2016. “Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection.” *Molecular Biology and Evolution* 33 (10): 2555–64. https://doi.org/10.1093/molbev/msw127.
Logsdon, Glennis A., Mitchell R. Vollger, and Evan E. Eichler. 2020. “Long-Read Human Genome Sequencing and Its Applications.” *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-0236-x.
Maller, Julian B., Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna M.M. Howson, et al. 2012. “Bayesian Refinement of Association Signals for 14 Loci in 3 Common Diseases.” *Nature Genetics* 44 (12): 1294–1301. https://doi.org/10.1038/ng.2435.
Mansournia, Mohammad Ali, Angelika Geroldinger, Sander Greenland, and Georg Heinze. 2018. “Separation in Logistic Regression: Causes, Consequences, and Control.” *American Journal of Epidemiology* 187 (4): 864–70. https://doi.org/10.1093/aje/kwx299.
Miretti, Marcos M., Emily C. Walsh, Xiayi Ke, Marcos Delgado, Mark Griffiths, Sarah Hunt, Jonathan Morrison, et al. 2005. “A High-Resolution Linkage-Disequilibrium Map of the Human Major Histocompatibility Complex and First Generation of Tag Single-Nucleotide Polymorphisms.” *American Journal of Human Genetics* 76 (4): 634–46. https://doi.org/10.1086/429393.
Nordin, Jessika, Adam Ameur, Kerstin Lindblad-Toh, Ulf Gyllensten, and Jennifer R.S. Meadows. 2020. “SweHLA: The High Confidence HLA Typing Bio-Resource Drawn from 1000 Swedish Genomes.” *European Journal of Human Genetics* 28 (5): 627–35. https://doi.org/10.1038/s41431-019-0559-2.
Palamara, Pier Francesco, Jonathan Terhorst, Yun S. Song, and Alkes L. Price. 2018. “High-Throughput Inference of Pairwise Coalescence Times Identifies Signals of Selection and Enriched Disease Heritability.” *Nature Genetics* 50 (9): 1311–17. https://doi.org/10.1038/s41588-018-0177-x.
Parham, Peter. 2005. “MHC Class I Molecules and KIRS in Human History, Health and Survival.” *Nature Reviews Immunology* 5 (3): 201–14. https://doi.org/10.1038/nri1570.
Petersdorf, Effie W., Theodore A. Gooley, Mari Malkki, Andrea P. Bacigalupo, Anne Cesbron, Ernette Du Toit, Gerhard Ehninger, et al. 2014. “HLA-C Expression Levels Define Permissible Mismatches in Hematopoietic Cell Transplantation.” *Blood* 124 (26): 3996–4003. https://doi.org/10.1182/blood-2014-09-599969.
Rahman, M. Shafiqur, and Mahbuba Sultana. 2017. “Performance of Firth-and LogF-Type

Penalized Methods in Risk Prediction for Small or Sparse Binary Data." *BMC Medical Research Methodology* 17 (1): 1–15. https://doi.org/10.1186/s12874-017-0313-9.
Robinson, James, Dominic J. Barker, Xenia Georgiou, Michael A. Cooper, Paul Flicek, and Steven G.E. Marsh. 2020. "IPD-IMGT/HLA Database." *Nucleic Acids Research* 48 (D1): D948–55. https://doi.org/10.1093/nar/gkz950.
Sabeti, Pardis C., Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, et al. 2007. "Genome-Wide Detection and Characterization of Positive Selection in Human Populations." *Nature* 449 (7164): 913–18. https://doi.org/10.1038/nature06250.
Sayers, Eric W., Richa Agarwala, Evan E. Bolton, J. Rodney Brister, Kathi Canese, Karen Clark, Ryan Connor, et al. 2019. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 47 (D1): D23–28. https://doi.org/10.1093/nar/gky1069.
Shieh, Mengkai, Tristan J. Hayeck, Anh Dinh, Jamie L. Duke, Nilesh Chitnis, Timothy Mosbruger, Ryan P. Morlen, et al. 2020. "Complex Linkage Disequilibrium Effects in HLA-DPB1 Expression and Molecular Mismatch Analyses of Transplantation Outcomes." *Transplantation* Publish Ah (3): 637–47. https://doi.org/10.1097/tp.0000000000003272.
Siewert, Katherine M., and Benjamin F. Voight. 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, Katherine M, and Benjamin F Voight. 2020. "BetaScan2: Standardized Statistics to Detect Balancing Selection Utilizing Substitution Data." *Genome Biology and Evolution* 12 (2): 3873–77. https://doi.org/10.1093/gbe/evaa013.
Slatkin, Montgomery. 2008. "Linkage Disequilibrium - Understanding the Evolutionary Past and Mapping the Medical Future." *Nature Reviews Genetics* 9 (6): 477–85. https://doi.org/10.1038/nrg2361.
Solovieff, Nadia, Jacqueline N. Milton, Stephen W. Hartley, Richard Sherva, Paola Sebastiani, Daniel A. Dworkis, Elizabeth S. Klings, et al. 2010. "Fetal Hemoglobin in Sickle Cell Anemia: Genome-Wide Association Studies Suggest a Regulatory Region in the 5′ Olfactory Receptor Gene Cluster." *Blood* 115 (9): 1815–22. https://doi.org/10.1182/blood-2009-08-239517.
Tajima, F. 1989. "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." *Genetics* 123 (3): 585–95. https://doi.org/PMC1203831.
Tanner, Julie Anne, and Rachel F. Tyndale. 2017. "Variation in CYP2A6 Activity and Personalized Medicine." *Journal of Personalized Medicine* 7 (4): 1–29. https://doi.org/10.3390/jpm7040018.
Tennessen, Jacob A., and Manoj T. Duraisingh. 2021. "Three Signatures of Adaptive Polymorphism Exemplified by Malaria-Associated Genes." *Molecular Biology and Evolution* 38 (4): 1356–71. https://doi.org/10.1093/molbev/msaa294.
Thul, Peter J., Lovisa Akesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, et al. 2017. "A Subcellular Map of the Human Proteome." *Science* 356 (6340). https://doi.org/10.1126/science.aal3321.
Tishkoff, S. A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos, G. Destro-Bisol, A. Drousiotou, et al. 2001. "Haplotype Diversity and Linkage Disequilibrium at Human G6PD: Recent Origin of Alleles That Confer Malarial Resistance." *Science* 293 (5529): 455–62. https://doi.org/10.1126/science.1061573.
Tripal, Philipp, Michael Bauer, Elisabeth Naschberger, Thomas Mörtinger, Christine Hohenadl, Emmanuelle Cornali, Mathias Thurau, and Michael Stürzl. 2007. "Unique Features of Different Members of the Human Guanylate-Binding Protein Family." *Journal of Interferon and Cytokine Research* 27 (1): 44–52. https://doi.org/10.1089/jir.2007.0086.
Uhlén, Mathias, Linn Fagerberg, Bjö M. Hallström, Cecilia Lindskog, Per Oksvold, Adil

Mardinoglu, Åsa Sivertsson, et al. 2015. “Tissue-Based Map of the Human Proteome.” *Science* 347 (6220). https://doi.org/10.1126/science.1260419.
Wakefield, Jon. 2007. “A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies.” *American Journal of Human Genetics* 81 (2): 208–27. https://doi.org/10.1086/519024.
———. 2009. “Bayes Factors for Genome-Wide Association Studies: Comparison with P -Values.” *Genetic Epidemiology* 33 (1): 79–86. https://doi.org/10.1002/gepi.20359.
Wiebe, Chris, Vasilis Kosmoliaptsis, Denish Pochinco, Craig J. Taylor, and Peter Nickerson. 2018. “A Comparison of HLA Molecular Mismatch Methods to Determine HLA Immunogenicity.” *Transplantation* 102 (8): 1338–43. https://doi.org/10.1097/TP.0000000000002117.
Wright, Stephen I., and Brian Charlesworth. 2004. “The HKA Test Revisited: A Maximum-Likelihood-Ratio Test of the Standard Neutral Model.” *Genetics* 168 (2): 1071–76. https://doi.org/10.1534/genetics.104.026500.