A Machine Learning Framework to Identify Selected Variants in Regions of Recent Adaptation

Andrew Jin

under the direction of
Mr. Joseph Vitti, Dr. Daniel Park, and Professor Pardis Sabeti
Harvard University

Research Science Institute
July 30, 2014

Abstract

Advances in DNA sequencing have enabled researchers to search for signatures of natural selection through genome-wide statistical scans, but the detected loci encompass thousands of candidate mutations. In this study, we develop a comprehensive machine learning approach that can distinguish the exact selected variants in regions of recent positive selection. The resulting models, after undergoing supervised learning on simulated population data, successfully classify 95 to 97 percent of selected SNPs at false positive rates of 2 to 3 percent. When applied to empirical full genome sequences from European, East Asian, and West African individuals from the 1000 Genomes Project, our methodology localizes signals to an average of 8 SNPs per region and identifies numerous variants in genes related to immune response, metabolic processes, sensory perception, and nervous system development.

Summary

Evolutionary forces play a central role in driving progress, yet very little is known about their precise mechanisms – namely the underlying causal alleles and the conferred advantageous traits. This study was the first to develop a machine learning approach to pinpoint mutations under positive selection. By linking our models' predicted candidate mutations to their relevant adaptive traits with biological experimentation, researchers can understand why our bodies are the way they are, examine our ancestors responses to selective pressures, and illuminate mechanisms through which we can adapt to modern-day challenges, such as infectious disease, climate change, and diet alteration.

1 Introduction

Darwin and Wallace first formulated the theory of positive natural selection in 1858, postulating that advantageous alleles enhance an individual's chances of reproduction and hence increase in prevalence in a population's gene pool over successive generations [1]. For billions of years, selective forces have played an integral part in driving progress on all tiers of biological organization, largely explaining the development of diverse, complex species from the most primitive of life forms. Exploring targets of more recent positive selection, especially in humans, not only allows for a better understanding of our ancestors' responses to selective pressures, such as diet alteration and climate change, but also offers insight into the various infectious diseases that continue to afflict modern-day society.

Despite positive selection's great import and foundational role in organismic biology, relatively little is known about the underlying mechanisms of adaptation, namely the specific mutations under selection and the resultant adaptive characteristics [2, 3]. Since fitness and advantageous traits are manifested on the phenotypic level, much of previous research has been directed at observing beneficial phenotypes (*e.g.* the different beak structures of Darwin's finches) and subsequently searching for the associated causal variant and evidence of selection [3, 4]. The extent of such knowledge, however, is highly limited, and aside from the clear examples of skin pigmentation (*SLC24A5*) [5], lactose tolerance (*LCT*) [6], malaria resistance (*HBB*) [7, 8], and high-altitude tolerance (*EPAS1*) [9], the vast majority of adaptive traits, such as immune system or metabolic changes, are exceedingly difficult to discern. Further, identifying the exact mutation responsible for the observed difference is even more of a challenge, since there are many candidate variants in each region of positive selection.

The recent advent of full genome sequencing presents an avenue to address the two aforementioned problems by enabling a shift in paradigm and paving the way for the reverse approach. Researchers can now utilize population genetic evidence to conduct exhaustive,

hypothesis-generating studies and identify potential selected variants for phenotype and functional characterization [3]. Such analyses rely on searching for three main types of signatures in DNA sequence: (i) Population differentiation – different regions experience unique selective pressures, so significant differences in allele frequency between geographically separate populations may signal positive selection. (ii) High-frequency derived (non-ancestral) alleles – the variant under selection brings nearby derived alleles to high frequency (the "hitchhiking"effect). (iii) Long haplotypes – in cases of positive selection, the beneficial allele rapidly rises in prevalence, reducing time for recombination to break down the selected variant's associations with neighboring variants [2, 4, 10].

Although statistical tests based on these three signatures have succeeded in detecting hundreds of loci potentially under positive selection, the hypothesized regions typically span hundreds of kilobases and encompass thousands of mutations [2]. Thus, new methods are required to pinpoint the exact causal variant, or at least reduce candidates to a tractable list for feasible functional characterization with biological experiments. The composite of multiple signals (CMS) test, developed by Grossman *et al.* in 2010, currently serves as the field standard for this task. CMS combines five population statistics (iHS, ΔiHH, and $XP\text{-}EHH$ for long haplotype signals; ΔDAF for high-frequency derived alleles; and F_{ST} for population differentiation) to calculate the posterior probability that a given single nucleotide polymorphism (SNP) has been targeted by positive selection [2]. By ensuring that a SNP simultaneously exhibits all three patterns of sequence variation, CMS substantially reduces false positives, while at the same time increasing sensitivity for the causal variant. So far the methodology has successfully discovered and elucidated the evolutionary role of a non-synonymous mutation in the Toll-like receptor 5 ($TLR5$) gene, which leads to altered NFκ-B signaling in response to bacterial flagellin [3].

However, CMS has shortcomings, as it relies extensively on approximate demographic data and population genetic models, which are time-consuming and labor-intensive to fur-

nish. Additionally, it functions as a naïve Bayesian classifier and assumes that its five statistics are completely independent, but three of the component scores test for long haplotypes and are clearly correlated. Therefore, through integration of various supervised classification methods, we developed a novel machine learning approach to identify selected SNPs in regions of recent adaptation. Our methodology involves computing hundreds of diverse input features and then training a model with data from coalescent simulations. The resulting framework, when applied to empirical full genome sequences from the 1000 Genomes Project, proves to be highly efficient, accurate, and generalizable (Figure 1).

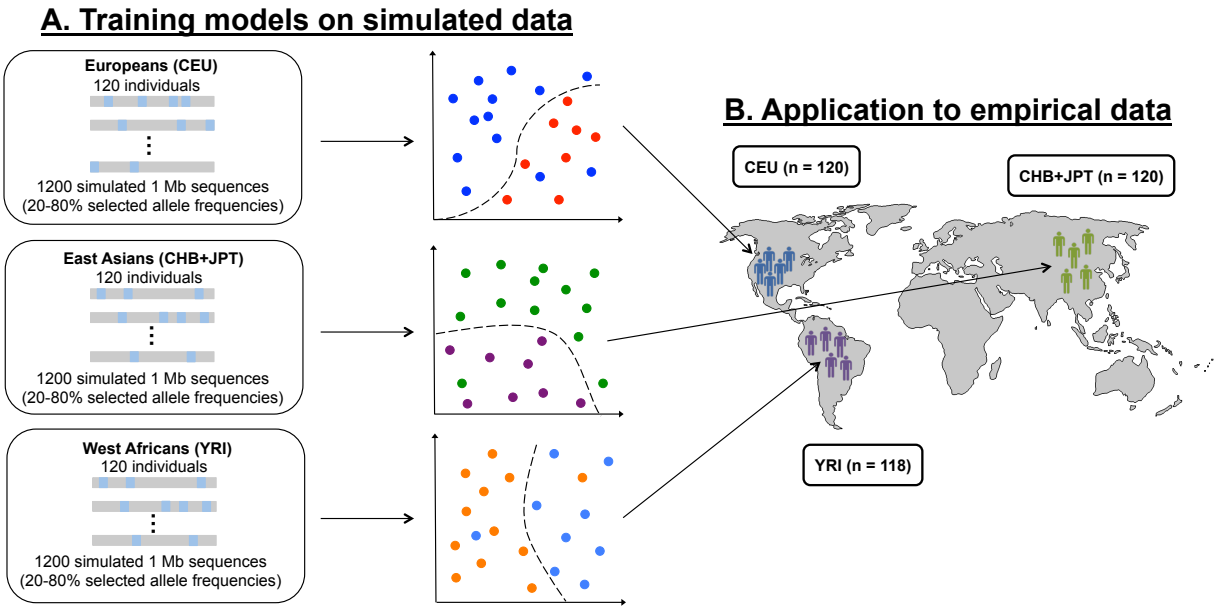

Figure 1: Graphical overview of the project workflow. (A) We trained three machine learning models on simulated DNA sequences, one for each of three populations (Europeans, East Asians, and West Africans). (B) We then applied the corresponding population classifier to identify selected SNPs in empirical data from 120 European, 120 East Asian, and 118 West African individuals from the 1000 Genomes Project.

2 Materials and Methods

2.1 Simulating regions under positive natural selection

Through standard coalescent simulation approaches with the cosi method [11], we generated calibrated genetic models of European, East Asian, and West African populations. Across the three populations, selected variants were set to reach different present-day frequencies (20, 40, 60, and 80 percent), creating a total of 12 datasets, each consisting of 300 replicates of 120 artificial chromosomes. Following the methodology of Grossman *et al.* [2, 3], such a chromosome consists of 1 Mb of simulated genetic sequence data and encompasses approximately 10,000 SNPs; we also assume that exactly one variant under positive selection appeared between 5,000 and 30,000 years ago, and resides exactly in the middle of the 1 Mb-long genomic region. Additional parameters, determined by the best-fit model of Schaffner *et al.* [11], control factors such as intensity of selection and migration rates.

2.2 Partitioning simulated data into training and validation sets

We pooled the data by population, so each of the three resulting datasets contained 1,200 selected variants and roughly 2,400,000 neutral variants. Five-sixths of the cases were randomly chosen to create a training set of 1,000 selected variants and about 2,000,000 neutral variants. All the remaining examples served as a validation set to assess classification accuracy.

2.3 Feature extraction from simulated DNA sequence data

For each SNP, we computed the five component statistical tests from CMS (F_{ST}, ΔDAF, iHS, ΔiHH, and $XP\text{-}EHH$) to utilize as input variables in our machine learning model. F_{ST} measures allele frequency differentiation between two populations and was computed with Weir and Cockerham's unbiased estimator [12]. The ΔDAF score assesses the prevalence

of high frequency derived alleles and is calculated by subtracting the average derived allele frequency in the non-selected populations from the derived allele frequency in the population under selection [2]. The iHS [13], ΔiHH [2], and $XP\text{-}EHH$ [14] statistics all derive from extended haplotype homozygosity (EHH), the probability that a haplotype, extending from the SNP for a certain genetic distance, is identical for two random chromosomes in the selected population.

We also implemented two additional test statistics – Tajima's D and locus specific branch lengths (LSBL). Tajima's D assesses high frequency derived alleles and differentiates between genomic regions evolving under neutrality and selection by comparing the number of segregating sites and the number of pairwise differences in a population [15]. LSBL offers an advantage over F_{ST} calculations, as it can compare allele frequency disparities between more than two populations [16].

In addition to the seven statistical tests discussed above, we also developed four types of novel higher-level features to capture information regarding the distribution of scores across an entire genomic region. First, for a given statistical test, a SNP's percentile rank compares its score to the scores of other SNPs in the candidate region. Second, scores usually increase in areas close to the selected variant, so for each SNP, we identified its 20 closest neighbors according to genetic distance. Computing mean score of the neighbors and score difference (how much a score is greater than its neighbors' mean score) provides insight into the various peaks of the score distribution. However, since score magnitudes may vary substantially between different replicates, we also introduced non-parametric versions (*e.g.* mean percentile rank of neighbors, percentile rank fold change, and percentile rank difference) of the previously discussed peak features. Third, we considered a SNP's genetic, physical, and Euclidean distance to the highest-scoring SNP in the candidate region. Finally, the relationships between statistics may be informative, since scores tend to exhibit weak correlation for neutral variants, but strong correlation in regions close to the selected variant. These

high-level features were extracted from each of the seven statistical scores, generating a total of 102 features for each SNP (Figure 2).

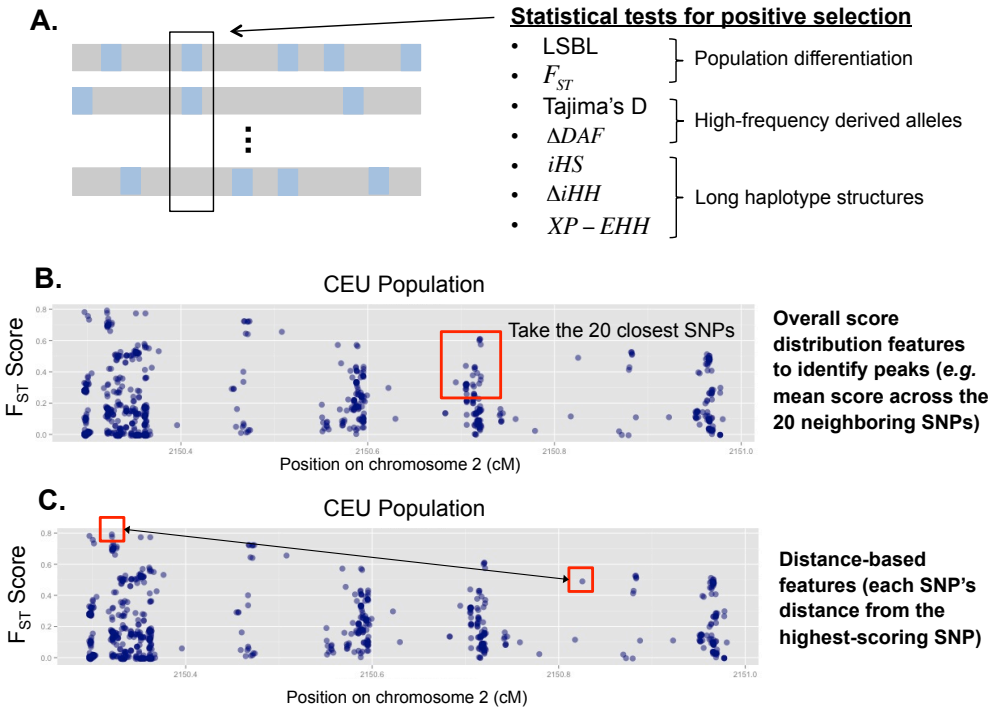

Figure 2: Extracting SNP features from sequence data. (A) For each SNP, we computed seven statistical scores testing for population differentiation, high-frequency derived alleles, and long haplotype structures. (B) To test if a SNP resides in a peak or high-scoring region, we compared its score to the scores of its 20 closest neighbors and calculated features such as mean score and mean percentile rank. (C) For all seven statistics, we calculated each SNP's distance from the highest-scoring SNP (measured by physical position, genetic distance, and Euclidean distance).

2.4 Sample selection to improve classifier performance

The data presented two major challenges, namely a high degree of class imbalance (for every selected variant, there are between 2,000 and 2,800 neutral variants) and the massive number of training examples (approximately 2,000,000 SNPs in each of the three population datasets). Developing supervised learning models on such data is computationally expensive

and exceedingly time consuming. Furthermore, the extreme imbalance would inevitably result in trivial classifiers that predict all SNPs to be neutral, achieving overall accuracy rates greater than 99.95 percent, but failing to identify true positives. The two most common and straightforward solutions are the non-heuristic random over-sampling and under-sampling approaches, which aim to achieve class balance through arbitrary duplication of minority class examples and removal of majority class examples, respectively. However, exact duplication of minority cases introduces no new information and may lead to overfitting, while random under-sampling potentially eliminates important data points that are essential to the learning process. Therefore, to address the two aforementioned challenges, we devised a novel three-step sample selection pipeline, tailored to the unique properties of our large, skewed training sets (Figure 3).

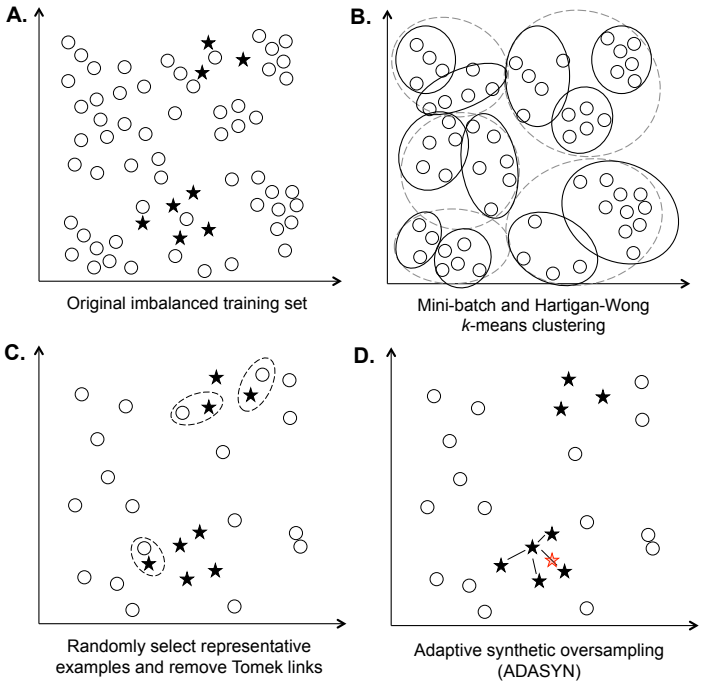

Figure 3: This figure offers a simple graphical representation of our three-step sample selection pipeline. Note that the example has 2 dimensions, while our datasets are in 102-dimensional space.

2.4.1 Clustering analysis to eliminate redundant examples

Following a concept introduced by Reinartz *et al.* [17], we proposed utilizing k-means clustering of majority class examples to identify groups of neutral SNPs that share similar feature values. In this way, only the representative cases are retained, allowing redundant examples to be removed without the loss of important information. Due to the large dataset size, standard k-means was too computationally expensive, so we first generated 50 rough clusters by applying the mini-batch k-means algorithm, which operates on small subsets of data during each iteration and updates cluster centroids through gradient descent [18]. The 50 clusters were further refined into a total of 2,500 sub clusters with a standard k-means implementation, and N points were chosen randomly from each sub cluster to serve as representative examples, with N selected according to the desired training set balance.

2.4.2 Removing Tomek links

The second step establishes well-defined class clusters by identifying Tomek links to eliminate noisy and borderline examples, thereby reducing the likelihood of overfitting during supervised learning and improving the models generalization capabilities. If $d(E_i, E_j)$ is the Euclidean distance between examples E_i and E_j, then the two examples form a Tomek link if they belong to different classes, and there is no example E_k such that $d(E_i, E_k) < d(E_i, E_j)$ or $d(E_j, E_k) < d(E_i, E_j)$ [19].

2.4.3 Synthetic over-sampling of minority examples

The final step involved a modified version of the Synthetic Minority Over-sampling Technique (SMOTE), which interpolates between each minority class example and its nearest neighbors to artificially generate new minority cases [20]. We utilized the adaptive synthetic sampling (ADASYN) method to assign weights and create varying amounts of artificial data for examples, depending on how challenging they are for the model to learn [21].

2.5 Developing supervised learning models

Missing inputs were imputed with average values, and data were normalized so that each feature had a mean of zero and variance of one across the training and validation examples. For each of our three datasets, we trained four different population-specific supervised learning models in the R programming environment (support vector machines with the e1071 package, random forests with the randomForest package, artificial neural networks with the nnet package, and LASSO regression models with the glmnet package). We tested neural network structures with different numbers of hidden neurons, and performed a grid search with varying and γ and cost (C) values to tune the support vector machines, as these models are especially sensitive to small parameter changes. After assessing classification accuracy with the held-out validation sets, we applied the trained classifiers to empirical data from the 1000 Genomes Project pilot phase [22], predicting selected SNPs from the full genome sequences of 120 Northern Europeans from Utah, 120 East Asians from China and Japan, and 118 West Africans from Nigeria.

2.6 Enrichment analysis to elucidate pathways under selection

We used the National Institute of Healths Database for Annotation, Visualization, and Integrated Discovery (DAVID) to discover pathways, themes, and functionally related gene clusters that are highly represented in the models' predictions. Through evaluation of enriched biological terms and pathways compiled from the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases, we were able to explore mechanisms that potentially explain responses to recent selective pressures.

3 Results

3.1 Assessing classifier performance

For a given genomic region displaying signals of positive selection, classifiers are tasked with pinpointing one or two causal variants in a background of thousands of neutral variants. Since they must identify small sets of potential SNPs that are tractable enough for experimental functional characterization, while ensuring with high confidence that the true selected variant is among the predicted candidates, we put particular emphasis on two performance measures: sensitivity, the proportion of selected SNPs that are correctly classified, and false positive rate (FPR), the proportion of neutral SNPs that are incorrectly classified as selected.

The optimal training set composition was determined to be a 5:1 ratio of negative to positive examples (10,000 neutral SNPs and 2,000 selected SNPs in each of the three population-specific training sets), and following supervised learning and parameter tuning, our machine learning models achieved high accuracy rates when classifying new, unseen cases from the validation sets. For example, the support vector machines, with $\gamma = 2^9$ and $C = 16$, on average correctly predicted 95.0 percent of the selected SNPs with a FPR of 2.6 percent. The average FPR is slightly inflated by the performance of the YRI model; positive selection is more difficult to discern in West African populations due to both greater SNP density and less pronounced linkage disequilibrium. Random forests with 500 trees exhibited similar performance, identifying 95.2 percent of causal variants at a 1.4 percent FPR. Neural networks (30 hidden nodes) and LASSO regression models achieved average sensitivities of 86.2 percent and 93.5 percent, and FPRs of 2.1 percent and 2.4 percent, respectively.

3.2 Developing an ensemble learning model

Each classifier had its strengths and weaknesses regarding prediction accuracy for different populations and selected allele frequencies. When compared to random forests, the support

vector machine model performed worse on the East Asian population, but displayed higher sensitivity for selected variants in the European population. Moreover, although random forests return substantially fewer false positives, their decision thresholds may be too stringent for application to empirical sequence data. Thus, we utilized a majority voting-based ensemble learning approach to improve prediction accuracy by integrating the outputs of the various individual classifiers, enabling them to compensate for one another's shortcomings. Three ensemble learners, one for each population, were trained and validated with the corresponding simulated datasets (Figure 4). For a more robust evaluation of classification performance on imbalanced data, we calculated the area under the receiver operating characteristic curve (AUROC), which is determined by plotting true positive rate against false positive rate at varying decision thresholds. A perfect classifier has an AUROC of 1, and any value between 0.9 and 1 is considered to be "excellent"[23]. All AUROC values for our ensemble models were greater than 0.99.

A.

	Actual Classes: Neutral	Actual Classes: Selected
Predictions: CEU Model		
Neutral	378,898	6
Selected	8,138	194

Neutral: 387,036 / # Selected: 200

B.

	Actual Classes: Neutral	Actual Classes: Selected
Predictions: CHB + JPT Model		
Neutral	356,866	10
Selected	7,368	190

Neutral: 364,234 / # Selected: 200

C.

	Actual Classes: Neutral	Actual Classes: Selected
Predictions: YRI Model		
Neutral	484,938	8
Selected	15,803	192

Neutral: 500,741 / Selected: 200

D.

Population	Accuracy	Sensitivity	FPR	AUROC
CEU	97.9%	97.0%	2.1%	0.994
CHB + JPT	98.0%	95.0%	2.0%	0.995
YRI	96.8%	96.0%	3.1%	0.991

Figure 4: (A-C) These contingency tables visualize the performance of the European (CEU), East Asian (CHB + JPT), and West African (YRI) ensemble classifiers. The columns represent instances in an actual class, while rows represent instances in a predicted class. (D) This table summarizes model performance with four measures: overall accuracy, sensitivity, false positive rate (FPR), and area under the receiver operating characteristic curve (AUROC).

3.3 Application to empirical data and positive controls

When applied to empirical full genome sequence data from European, East Asian, and West African individuals in the 1000 Genomes Project, the ensemble classifiers effectively localized signals of positive selection and pinpointed causal variants in candidate genomic regions. In the CEU and YRI populations, we narrowed down the number of potential SNPs from about 1,500 per region to an average of 10. The CHB + JPT model's localization ability was even more substantial. For each region of positive selection, an average of 4 variants were predicted to be selected out of a background of more than 1,550.

Although very few adaptive traits and causal variants are currently known, several selected SNPs have been identified and rigorously characterized, namely mutations in or around the *LCT*, *PCDH15*, *SLC25A4*, and *EDAR* genes. As further validation of our methodology, we utilized these genomic regions as positive controls, and indeed the models correctly pinpointed the selected SNP in each case. For instance, the rs182549 polymorphism, located on the *MCM6* gene in chromosome 2, influences the lactase (*LCT*) gene and is linked to lactase persistence in European populations. When applied to the 951 SNPs in the candidate region, the CEU model identified 20 as selected, assigning the second highest confidence score to the rs182549 polymorphism.

The *PCDH15* gene plays a crucial role in sensory perception and is associated with retinal photoreceptor maintenance, inner ear hair cell development, Usher Syndrome and hearing loss [2]. The nonsynonymous D440A mutation (rs4935502) and a polymorphism in the gene's transcription factor binding site (rs16905686) have been hypothesized to be targets of positive selection in East Asian populations. The CHB + JPT classifier successfully localized the signal and narrowed down the candidates from 2,584 SNPs to 7, assigning rs4935502 and rs16905686 with the fourth and sixth highest confidence scores. In comparison to the field standard CMS test, our methodology performs similarly on the LCT case, but offers significant improvements when detecting the two *PCDH15* variants (Figure 5).

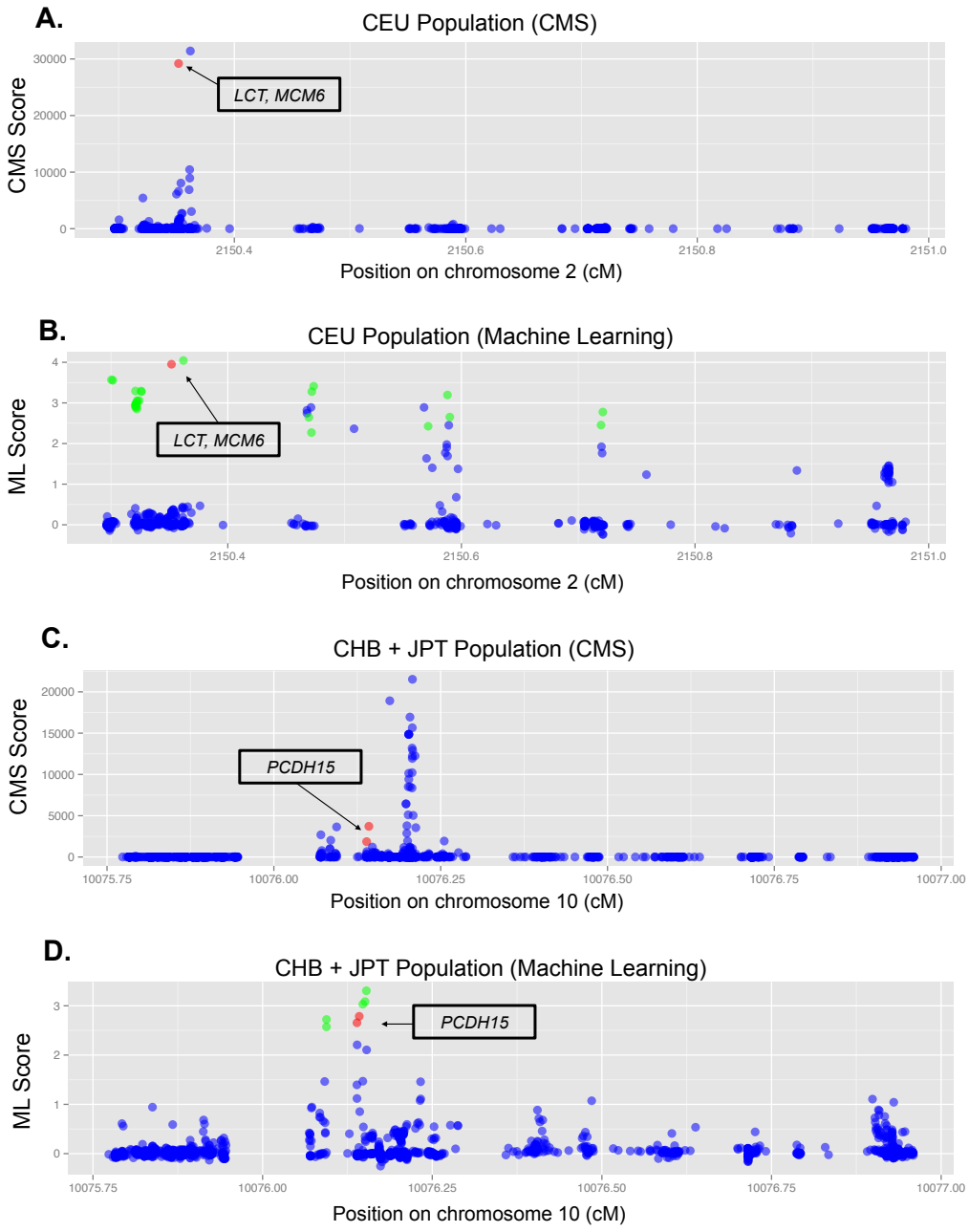

Figure 5: (A) CMS scores for SNPs in the candidate region containing the selected lactase persistence allele. The red dot represents the actual causal variant. (B) Machine learning decision values in the same LCT candidate region. The red dot represents the actual causal variant, green dots are SNPs identified as selected, and blue dots are SNPs predicted to be neutral. (C-D) CMS and machine learning scores in the candidate region containing the selected *PCDH15* variants. Our machine learning methodology performs significantly better than CMS in this case.

In addition to the four well-characterized positive control variants, our machine learning models detected SNPs in various genes proposed to be under positive selection and implicated in important biological processes. A few examples include *FOXP1* in East Asians, as well as *FOXP2*, *LARGE*, and *HBB* in West Africans. *FOXP1* and *FOXP2* play important roles in development of vital organs and speech capabilities [24], while mutations in LARGE and HBB confer resistance to Lassa fever and malaria, respectively [25, 8].

3.4 Computational functional annotation

We compiled a list of all genes within 50 kilobases of SNPs that were predicted to be under positive selection and conducted gene set enrichment analysis to gain a better understanding of the affected biological pathways and potential adaptive phenotypes. Consistent with our understanding of recent selective pressures such as infectious disease, climate change, and diet alteration, numerous gene sets related to immune response, metabolic processes, sensory perception, and nervous system development were observed to be highly represented across all three populations (Figure 6). Some population-specific results include enrichment of skin pigmentation-related genes in the European population and oxygen transport-related genes in the West African population.

3.5 Discovering novel targets of selection

After confirming the effectiveness of our approach on empirical sequences – by validating with multiple positive controls – we applied the classifiers to the 1000 Genomes Project data to discover new variants that have not been previously studied as targets of natural selection (Figure 7). Out of the 3077 SNPs classified as selected, 20 were found to be nonsynonymous mutations that changed the amino acid sequence of the protein. Functional annotation of these hits, 18 of which were novel, reveals important biological processes (Table 1).

	CEU	P-value	CHB+JPT	P-value	YRI	P-value
Immune response	CD4 T cell differentiation	0.017	B cell differentiation	**0.0063**	MHC peptide loading complex	**1.6E-7**
	Positive regulation of B cell proliferation	0.018	Lymphocyte activation	0.089	Antigen processing and presentation	**4.1E-4**
Metabolism	Fat-soluble vitamin metabolic process	0.014	Vitamin metabolic process	**1.5E-4**	Lipoprotein metabolic process	0.079
	Vitamin A metabolic process	0.027	Fatty acid metabolism	**0.0042**	Cellular polysaccharide metabolic process	0.14
Sensory perception and brain	Neural crest cell differentiation	0.014	CNS neuron development	0.087	Sensory perception of chemical stimulus	**1.2E-4**
	Neural crest cell development	0.014	Sensory perception of light stimulus	0.43	Olfaction	**4.1E-4**

Figure 6: Gene sets related to immune response (*e.g.* CD4 T cell differentiation), metabolic processes (*e.g.* fatty acid metabolism), and sensory perception and brain functions (*e.g.* olfaction) are highly enriched for selection in all three populations. This is consistent with recent selective pressures, such as infectious disease, climate change, and diet alteration.

Chrom	Population	Position	SNP ID	Gene	Mutation	Potential Biological Processes
1	YRI	171784170	rs7551131	SLC9C2	T481M	Pneumonia, neurological diseases
2	CEU	74543547	rs2268416	MOGS	P293S	Resistance to viral infection, N-linked oligosaccharide processing
2	CEU	162917139	rs17783344	GCA	S80A	Preeclampsia susceptibility, cardiovascular disease, type 1 diabetes
5	CEU	131704219	rs1050152	SLC22A4	L327F	Crohn's disease, type 1 diabetes, rheumatoid arthritis
9	CEU	129924574	rs13283456	PTGES2	R107H	Body mass index, type 2 diabetes in Germans
11	YRI	5320871	rs12273630	OR51B5	V154I	Olfactory receptor
11	YRI	5431658	rs10450603	OR51I2	R122C	Olfactory receptor
11	YRI	5466972	rs7935144	OR52D1	R154C	Olfactory receptor
11	YRI	5467151	rs7924754	OR52D1	D213E	Olfactory receptor
12	YRI	121907152	rs7972242	HIP1R	K504Q	Spindle attachment for chromosomal segregation during mitosis
16	CEU	76978276	rs12918952	WWOX	A179T	High-density lipoprotein cholesterol levels in French individuals
19	CEU	38297140	rs10416265	GPATCH1	H724R	Meningitis and osteoporosis susceptibility
19	CEU	38297152	rs10421769	GPATCH1	L728S	Meningitis and osteoporosis susceptibility

Table 1: This table shows potentially selected nonsynonymous mutations that have been associated with important biological processes or phenotypes.

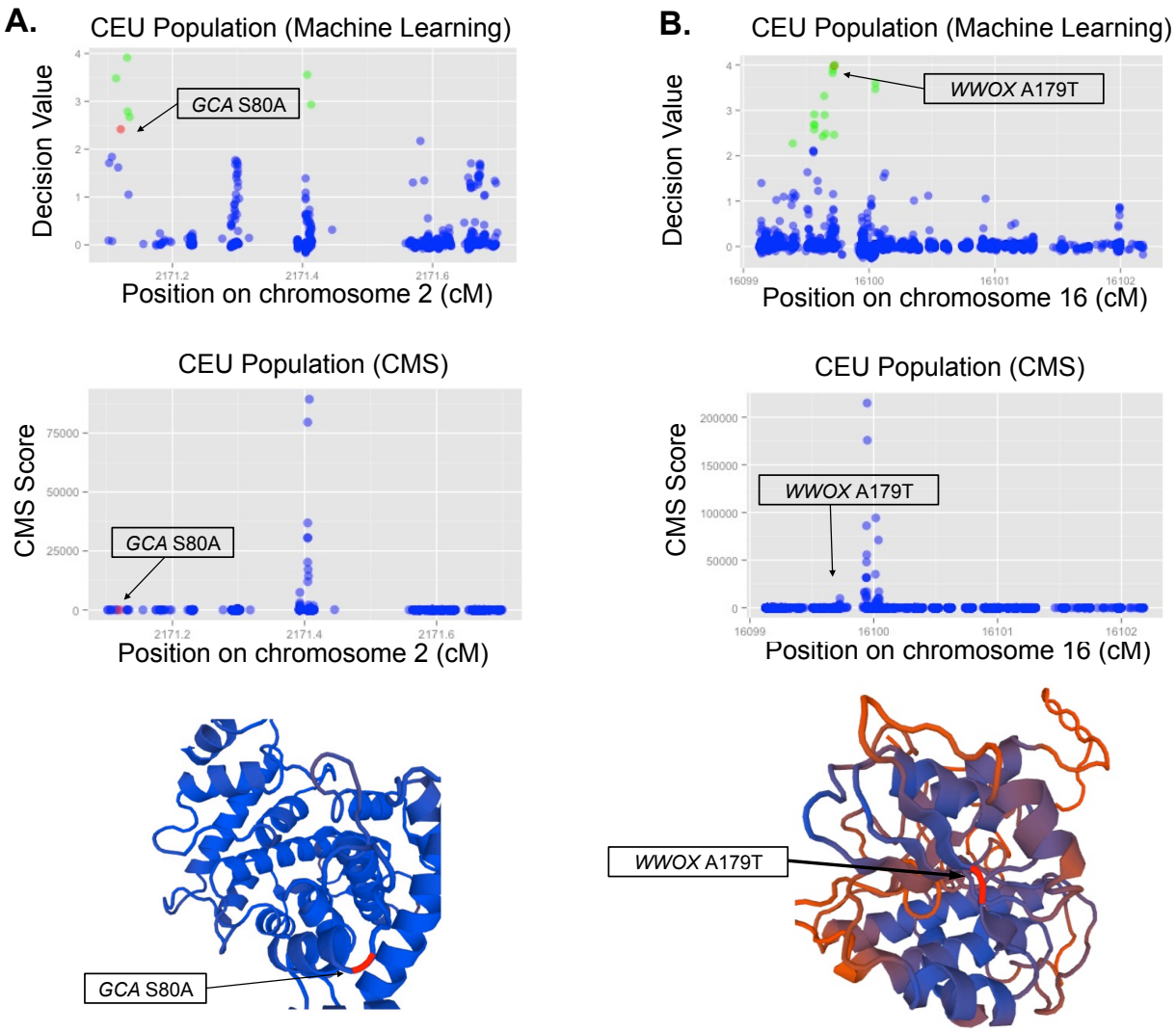

Figure 7: A) CMS and machine learning decision value plots for the region containing the *GCA* S80A nonsynonymous mutation. The putative causal allele is highlighted in red, and all other SNPs predicted to be selected are green. (B) CMS and machine learning decision value plots for the region containing the *WWOX* A179T nonsynonymous mutation. Homology models were generated and visualized with SWISS-MODEL, and the mutated residues are highlighted in red. Plots for the other predicted novel nonsynonymous mutations can be found in the appendix.

For example, the *PTGES2* Arg^{107}→His and the nonpolar-to-polar *WWOX* Ala^{179}→Thr mutations showed strong signals of positive selection in the European population and have been linked to body mass index, type 2 diabetes, and high-density lipoprotein cholesterol levels in German and French cohorts [26, 27, 28]. *WWOX* is of particular interest, as it resides in human accelerated region 6 and may therefore help explain how humans differentiated from primates [29]. Furthermore, numerous nonsynonymous variants predicted to be selected in the CEU dataset play crucial roles in disease susceptibility. A nonpolar-to-polar amino acid change (Leu^{327}→Phe) in *SLC22A4* is associated with Crohn's disease, type 1 diabetes, and rheumatoid arthritis [30, 31], while mutations in *MOGS*, which encodes the first enzyme in the N-linked oligosaccharide processing pathway, potentially confer resistance to viral infections, such as influenza, herpes, and hepatitis C [32]. In the West African population dataset, four potentially selected nonsynonymous mutations were pinpointed in various olfactory receptor genes on chromosome 5, and the Lys^{504}→Gln variant was classified as selected in the *HIP1R* gene, which regulates the function of spindle microtubules during chromosomal segregation [33].

3.6 Exploring potential selected variants in regulatory regions

Our models' predictions suggest that the majority of recent positive selection has been driven by mutations in noncoding regions of the genome, as only 20 of the 3077 candidate variants lead to amino acid changes. To investigate the remaining mutations, we downloaded genome-wide association study (GWAS) data from the UCSC Genome Browser database and identified relevant phenotypic traits for 14 regulatory SNPs (Table 2). Interestingly, multiple intronic SNPs were related to cholesterol and phospholipid levels, chronic kidney disease, and cardiac structure. These candidates may be analyzed through expression quantitative trait loci (eQTL) studies better understand how they lead to the observed traits.

Chrom	Position	SNP ID	Gene	Trait	Sample	P-value	Citation
2	135554376	rs7570971	RAB3GAP1	Cholesterol	94,595 Europeaan ancestry individuals	1.0E-13	Willer *et al.* 2013
15	73505722	rs8028182	SIN3A	Sudden cardiac arrest	89 European cases, 520 controls	3.0E-06	Aouizerat *et al.* 2011
16	76816311	rs2059238	WWOX	Cardiac structure and function	12,612 European ancestry individuals	3.0E-06	Vasan *et al.* 2009
11	61468051	rs2521572	BEST1	Phospholipid levels (plasma)	8,866 European ancestry individuals	2.0E-09	Lemaitre *et al.* 2011
17	61529007	rs2319125	CEP112	Glycoprotein levels (plasma)	306 European ancestry individuals	1.0E-06	Athanasiadis *et al.* 2013
15	43428517	rs2453533	GATM	Chronic kidney disease	67,093 European ancestry individuals	5.0E-22	Kottgen *et al.* 2010
15	43486085	rs2467853	SPATA5L1	Renal function and chronic kidney disease	2,388 European cases, 17,489 controls	6.0E-14	Kottgen *et al.* 2010
1	30400299	rs2180233	Intergenic	Attention deficit hyperactivity disorder	938 European ancestry trios	9.0E-06	Anney *et al.* 2008
17	61073004	rs8074751	CCDC46	Attention deficit hyperactivity disorder	735 trios from 732 families	1.0E-06	Mick *et al.* 2010

Table 2: This table shows potentially selected noncoding mutations that have been associated with important phenotypes through previous GWAS studies.

3.7 Generalizing to new populations

To identify variants under positive selection in a certain population from the empirical 1000 Genomes Project data, we first trained and validated an ensemble learning model with the corresponding simulated population data (*e.g.* a model trained on the simulated CEU data was used to make predictions for the empirical CEU sequence data). However, for many of the newly profiled populations from the recent 1000 Genomes Phase 3 release, such as Gujarati Indians and Gambians, demographic information and calibrated genetic models are either unavailable or highly inaccurate. We evaluated the generalization capabilities of our models to investigate the possibility of utilizing this study's methodology to detect selection in unstudied populations, and the results were quite promising.

When applied to a validation set of CHB + JPT simulated data, the ensemble model trained with the CEU simulated data classified selected SNPs with a sensitivity of 95.0 percent and FPR of 2.0 percent. It also successfully pinpointed the two *PCDH15* positive controls in the empirical 1000 Genomes Project DNA sequences, narrowing down the 2,584 candidates to a set of 9 variants. Similarly, the CHB + JPT ensemble model achieved 97.0 percent sensitivity and a 2.1 percent FPR when detecting selection in the simulated CEU data and pinpointed the LCT positive control to a set of 16 variants. Lastly, when applied to the CEU and CHB + JPT simulated validation sets, the classifier built with YRI data maintained high classification accuracy (99 and 96.5 percent sensitivity), although there

was a slight increase in FPR (3.9 and 3.6 percent). These results suggest the possibility of using our classifiers to analyze unstudied population data, thus circumventing the need for developing and validating new demographic models.

4 Discussion

Even with strong population genetic evidence for selection and sufficient support for biological impact through extensive computational functional annotation, a SNP cannot truly be considered a causal variant until the relevant adaptive trait is elucidated and characterized with *in vitro* or *in vivo* experimentation. Nevertheless, this study offers a major improvement in the most important step of the selection detection process. After conducting genome-wide scans to search for signatures of selective sweeps, one can utilize machine learning to evaluate thousands of potential candidates and pinpoint the signal to a small list of variants that is tractable enough for biological validation.

This study's methodology offers three main improvements over the CMS test, which currently serves as the field standard. First, the ensemble learning models provide increased localization abilities, decreasing the number of false positives while maintaining high sensitivity for the selected variant. CMS captures 90 percent of selected SNPs in the simulated data and when applied to the empirical 1000 Genomes Project data, returns a set of 20 to 100 candidate SNPs per region. Our classifiers correctly identify 95 to 97 percent of selected SNPs and return an average of 8 candidate SNPs per region. Second, CMS uses Bayesian calculations and hence relies extensively on probability distributions from simulated data, which not only are labor intensive and time consuming to generate, but also require a realistic demographic model for the population or species under analysis. It is impossible to effectively apply the test to unstudied populations, such as South Asians from the recent Phase 3 release of the 1000 Genomes Project, since demographic information is unavailable

or inaccurate. However, a machine learning framework built on a standard calibrated European, East Asian, or African population model can be applied to new populations without any additional training or simulations. Finally, while CMS often has difficulty detecting low-frequency selected variants, our machine learning model maintains sensitivities greater than 90 percent when identifying SNPs fixated at present-day frequencies of 20 and 40 percent.

5 Conclusion

Although the advent of DNA sequencing has revolutionized the field of evolutionary genetics by enabling researchers to leverage genome-wide scans that identify loci exhibiting signatures of positive natural selection, the detected loci are large and cover thousands of mutations. The current field standard, the composite of multiple signals (CMS) test, can localize the signal to a set of 20 to 100 candidate variants, but has three limitations. It assumes that different population genetics statistics are independent, depends on labor-intensive and time-consuming demographic simulations, and has difficulty distinguishing low-frequency variants. In order to address these challenges, this study was the first to develop machine learning models to pinpoint the exact causal mutations in regions of positive selection. When classifying full genome sequence data from the 1000 Genomes Project, the approach localizes signals to an average of 8 SNPs per region and correctly predicted well-known causal variants in the *LCT*, *PCDH15*, *SLC25A4*, and *EDAR* genes. Moreover, many of the SNPs identified as selected reside on or near genes that code for important processes, such as immune response, metabolism, and nervous system development. In the future, we plan to fully validate top candidate SNPs through rigorous *in vitro* and *in vivo* functional characterization experiments and apply the classifiers to unstudied populations to fully demonstrate the effectiveness of our predictive framework.

6 Acknowledgments

I would like to thank my mentors Mr. Joseph Vitti, Dr. Daniel Park, and Professor Pardis Sabeti from Harvard University, as well as my tutor Ms. Ana Lyons and Research Science Institute (RSI) alumni James Thomas, Dominik Rabiej, and Ava Chen, for their valuable advice and guidance. I would like to acknowledge Ilya Shlyakhter for generating simulation data and calculating CMS component scores. I would also like to thank RSI, my sponsors, the Center for Excellence in Education, the Broad Institute, and Massachusetts Institute of Technology for providing me with such a wonderful research opportunity this summer.

References

[1] C. R. Darwin and A. R. Wallace. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London*, 3(9):45–62, 1858.

[2] S. R. Grossman, I. Shylakhter, E. K. Karlsson, E. H. Byrne, S. Morales, G. Frieden, E. Hostetter, E. Angelino, M. Garber, O. Zuk, E. S. Lander, S. F. Schaffner, and P. C. Sabeti. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 327(5967):883–6, 2010.

[3] S. R. Grossman, K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, A. Yen, D. J. Park, D. Griesemer, E. K. Karlsson, S. H. Wong, M. Cabili, R. A. Adegbola, R. N. Bamezai, A. V. Hill, F. O. Vannberg, J. L. Rinn, . G. Project, E. S. Lander, S. F. Schaffner, and P. C. Sabeti. Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4):703–13, 2013.

[4] P. C. Sabeti, S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. Mikkelsen, D. Altshuler, and E. S. Lander. Positive natural selection in the human lineage. *Science*, 312(5780):1614–20, 2006.

[5] R. L. Lamason, M.-A. P. Mohideen, J. R. Mest, A. C. Wong, H. L. Norton, M. C. Aros, M. J. Jurynec, X. Mao, V. R. Humphreville, J. E. Humbert, S. Sinha, J. L. Moore, P. Jagadeeswaran, W. Zhao, G. Ning, I. Makalowska, P. M. McKeigue, D. O'Donnell, R. Kittles, E. J. Parra, N. J. Mangini, D. J. Grunwald, M. D. Shriver, V. A. Canfield, and K. C. Cheng. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, 310(5755):1782–6, 2005.

[6] T. Bersaglieri, P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, 74(6):1111–20, 2004.

[7] M. Currat, G. Trabuchet, D. Rees, P. Perrin, R. M. Harding, J. B. Clegg, A. Langaney, and L. Excoffier. Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the Beta(S) Senegal mutation. *American Journal of Human Genetics*, 70(1):207–23, 2002.

[8] J. Ohashi, I. Naka, J. Patarapotikul, H. Hananantachai, G. Brittenham, S. Looareesuwan, A. G. Clark, and K. Tokunaga. Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *American Journal of Human Genetics*, 74(6):1198–208, 2004.

[9] X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie,

H. Wu, M. Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, Asan, P. Ni, G. Tian, J. Xu, X. Liu, T. Jiang, R. Wu, G. Zhou, M. Tang, J. Qin, T. Wang, S. Feng, G. Li, Huasang, J. Luosang, W. Wang, F. Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, S. Li, H. Yang, R. Nielsen, J. Wang, and J. Wang. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329(5987):75–8, 2010.

[10] J. J. Vitti, S. R. Grossman, and P. C. Sabeti. Detecting natural selection in genomic data. *Annual Review of Genetics*, 47:97–120, 2013.

[11] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15(11):1576–83, 2005.

[12] B. S. Weir and W. G. Hill. Estimating f-statistics. *Annual Review of Genetics*, 36:721–50, 2002.

[13] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS Biology*, 4(3), 2006.

[14] P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander, and T. I. H. Consortium. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–8, 2007.

[15] F. Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3):585–95, 1989.

[16] M. D. Shriver, G. C. Kennedy, E. J. Parra, H. A. Lawson, V. Sonpar, J. Huang, J. M. Akey, and K. W. Jones. The genomic distribution of population substructure in four populations using 8,525 autosomal snps. *Human Genomics*, 1(4):274–86, 2004.

[17] T. Reinartz. A unifying view on instance selection. *Data Mining and Knowledge Discovery*, 6(2):191–210, 2002.

[18] D. Sculley. Web-scale k-means clustering. *Proceedings of the 19th International Conference of World Wide Web*, pages 1177–1178, 2010.

[19] I. Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, 6:769–772, 1976.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and P. W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[21] H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *IEEE Xplore*, pages 1322–1328, 2008.

[22] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[23] R. H. E. Khouli, K. J. Macura, P. B. Barker, M. R. Habba, M. A. Jacobs, and D. A. Bluemke. The relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced (dce) mri of the breast. *Journal of Magnetic Resonance Imaging*, 30(5):999–1004, 2009.

[24] W. Shu, M. M. Lu, Y. Zhang, P. W. Tucker, D. Zhou, and E. E. Morrisey. FOXP2 and FOXP1 cooperatively regulate lung and esophagus development. *Development*, 134:1991–2000, 2007.

[25] K. G. Andersen, I. Shylakhter, S. Tabrizi, S. R. Grossman, C. T. Happi, and P. C. Sabeti. Genome-wide scans provide evidence for positive selection of genes implicated in lassa fever. *Philosophical Transactions of the Royal Society of London*, 367(1590):868–77, 2012.

[26] A. Fischer, H. Grallert, M. Bohme, C. Gieger, I. Boomgaarden, H. Wichmann, F. Doring, and T. Illig. Association analysis between the prostaglandin E synthase 2 R298H polymorphism and body mass index in 8079 participants of the KORA study cohort. *Genetic Testing and Molecular Biomarkers*, 13(2):223–6, 2009.

[27] I. Nitz, E. Fisher, H. Grallert, Y. Li, C. Gieger, D. Rubin, H. Boeing, J. Spranger, I. Lindner, S. Schreiber, W. Rathmann, H. Gohlke, A. Doring, H. E. Wichmann, J. Schrezenmeir, F. Doring, and T. Illig. Association of prostaglandin E synthase 2 (PTGES2) Arg298His polymorphism with type 2 diabetes in two German study populations. *The Journal of Clinical Endocrinology and Metabolism*, 92(8):3183–8, 2007.

[28] Z. Dastani, P. Pajukanta, M. Marcil, N. Rudzicz, I. Ruel, S. D. Bailey, J. C. Lee, M. Lemire, J. Faith, J. Platko, T. J. H. John Rioux, D. Gaudet, J. C. Engert, and J. Genest. Fine mapping and association studies of a high-density lipoprotein cholesterol linkage region on chromosome 16 in French-Canadian subjects. *European Journal of Human Genetics*, 18(3):342–7, 2010.

[29] G. B. Kamm, F. Pisciottano, R. Kliger, and L. F. Franchini. The developmental brain gene NPAS3 contains the largest number of accelerated regulatory sequences in the human genome. *Molecular Biology and Evolution*, 30(5):1088–102, 2013.

[30] O. Hradsky, P. Dusatkova, M. Lenicek, J. Bronsky, D. Duricova, J. Nevoral, L. Vitek, M. Lukas, and O. Cinek. Two independent genetic factors responsible for the associations of the IBD5 locus with Crohn's disease in the Czech population. *Inflammatory Bowel Diseases*, 17(7):1523–9, 2011.

[31] J. L. Santiago, A. Martinez, H. de la Calle, M. Fernandez-Arquero, A. Figueredo, E. G. de la Concha, and E. Urcelay. Evidence for the association of the SLC22A4 and SLC22A5 genes with type 1 diabetes: a case control study. *BioMed Central Medical Genetics*, 7(54), 2006.

[32] M. A. Sadat, S. Moir, T.-W. Chun, P. Lusso, G. Kaplan, L. Wolfe, M. J. Memoli, M. He, H. Vega, L. J. Kim, Y. Huang, N. Hussein, E. Nievas, R. Mitchell, M. Garofalo, A. Louie, D. C. Ireland, C. Grunes, R. Cimbro, V. Patel, G. Holzapfel, D. Salahuddin, T. Bristol, D. Adams, B. E. Marciano, M. Hegde, Y. Li, K. R. Calvo, J. Stoddard, J. S. Justement, J. Jacques, D. A. L. Priel, D. Murray, P. Sun, D. B. Kuhns, C. F. Boerkoel, J. A. Chiorini, G. D. Pasquale, D. Verthelyi, and S. D. Rosenzweig. Glycosylation and hypogammaglobulinemia and and resistance to viral infections. *The New England Journal of Medicine*, 370(17):1615–25, 2014.

[33] S. Park. Huntingtin-interacting protein 1-related is required for accurate congression and segregation of chromosomes. *BMB Reports*, 43(12):795–800, 2010.