How do environmental changes influence human health? One major challenge lies in the complex multifactorial nature of these influences. My research addresses this by exploring how historical environmental changes have shaped the human genome. Specifically, my research focus on answering two questions. First, how **natural selection** affects **human** **infectious diseases**, particularly on how humans adapt to viruses. Using protein structures and functional genomics, I explore how human proteins and regulatory elements evolve in response to viruses. Second, how natural selection affects human **chronic diseases**. I aim to understand the natural selection on various non-coding genomic regions and its impact on chronic diseases. Here, I describe my prior work in each area and outline future plans for my independent lab.

**Previous work**

I developed a pipeline to detect enrichment of recent positive selection in a group of genes and a method to infer distribution of fitness effects of human noncoding mutations. Using these approaches, I identified a key mechanism by which human proteins adapt to viruses, make the first full distribution of fitness effects of mutations in human noncoding genomic regions, where most variants associated with chronic diseases reside, and proposed a conceptual model for the origin of Mendelian diseases, supported by empirical evidence.

***Protein stability changes as a major mechanism against viruses***

Viruses were a dominant driver of adaptation in thousands of human proteins that physically interact with viruses (VIPs for Virus-Interacting Proteins)1. Adaptative mutations have been found both at surface and interior throughout VIPs. However, well-studied cases of virus-driven adaptation in immune factors and viral receptors usually focus on limited surface regions of VIPs. How do other adaptive mutations might influence protein functions? Protein stability—the balance between folded and unfolded forms—can be influenced by widespread amino acid mutations and thus may be an important protein feature in virus-drive protein adaptation. Using protein stability changes predicted from AlphaFold 2 structures, I found that amino acids changes altering protein stability experienced significantly elevate adaptive evolution in VIPs comparing to other proteins. This suggests that the evolution of protein stability, and thus functional protein abundance, is a key mechanism of host adaptation to viruses. This work is in revision at *Molecular Biology and Evolution* (Di et al., 2022, biorxiv2).

***Fitness landscape of non-coding mutations in the human genome***

Most previous estimates of non-coding mutation fitness effects limited to a small genomic region, and a few categories of selective strength, highlighting the need for more complete estimates. I combined genomic annotations from epigenomic chromatin profiles across cell types with genetic variation

patterns and population genetic models to estimate the full distribution of fitness effects (DFE) of human noncoding mutations. My findings reveal DFEs of functional regions, like enhancers and promoters, and genomic regions under varying constraints. Notably, highly conserved sites—often highlighted in biomedical studies—capture only a small fraction of deleterious mutations; for instance, the top 5% of conserved non-coding sites capture <20% of deleterious mutations. This indicates that functional non-coding sites are more broadly distributed across evolutionary constraints than previously thought. Additionally, I observed dynamic evolution of gene regulation, evidenced by shifting selection coefficients over deep evolutionary time (Di et al., 2025, *bioRxiv*3)

***Lag of environmental changes may result in Mendelian diseases***

Failing to adapt to new environments may cause diseases. My previous work found that pre-existing deleterious mutations slow recent adaptation in human Mendelian disease genes, as evidenced by fewer selective sweeps compared to non-disease genes with similar levels of constraint and other factors influencing adaptation. Despite slow recent adaptation, disease genes show similar long-term protein adaptation rates to non-disease genes, as evidenced by similar proportions of adaptive amino acid changes over millions of years of human evolution. This suggests that disease genes do not have constitutively fewer adaptive mutations, but instead experience a transient inability to adapt as fast as the rest of the genome in recent times (*eLife*, Di et al., 20214).

**Future direction 1. History and Mechanisms of adapting against viruses**

Viruses are a dominant driver of protein adaptation1, but whether this is dominant by large-effect mutations or polygenic responses remains an open question. The role of noncoding elements is also unclear. My research aims to map virus-driven selection in human history and uncover the genetic mechanisms of adaptation, shedding light on the evolution of human immunity and health.

***When and how did humans adapt to viruses?***

Prior studies show that virus-interacting proteins (VIPs) experience elevated rates of strong adaptation, while rates of weak adaptation appear similar between VIPs and non-VIPs5. At the same time, growing evidence suggests that weak positive selection resulting in subtle frequency shifts, is common in recent human history6,7,8. This raises a key question: is adaptation in VIPs dominated by strong selective events, or weak, polygenic responses? I address this question by comparing positive selection in VIPs and non-VIPs using state-of-the-art methods that infer both the timing and strength of selection.

Both polygenic adaptation and selection on a few standing variants produce weak signals, as do responses to localized or recurring viral epidemics. Biologically, weakly beneficial mutations are less likely to fix, meaning approaches focused on fixed differences over long timescales (e.g., past million years5) may be biased toward detecting only strong adaptation. Moreover, because the dynamics of weakly selected variants closely resemble neutrality, weak selection is technically harder to detect. These challenges are compounded when weak and strong selection occur together, further obscuring signals.

To address these challenges, I will apply methods with improved power for detecting weak selection. These include Ancestral Recombination Graph (ARG)-based inference with **CLUES2**9 and machine learning approaches such as **TrIdent**10 and **Flex-Sweep**11. Unlike earlier methods that only yield summary statistics, CLUES2 and TrIdent also estimate the onset time of selection, providing finer resolution on when viral adaptation occurred. Using these tools, I will reconstruct the timing and strength of selection in VIPs. Viral exposure may intensify episodically with the rise of dense human settlements or increased contact with animals. Alternatively, humans experienced a constant viral pressure. With timing and selective strength, we could test which selective pressure is dominant and explore the key adaptive mode in different scenarios.

I will then compare the number of positively selected genes and the strength of selection in VIPs versus matched non-VIPs across historical intervals. Because demographic events (e.g., expansions, bottlenecks) can mimic or mask selection, we will control for such background factors to isolate VIP-specific signals of viral-driven adaptation by comparing VIPs to control non-VIPs and matched for other confounding factors. By comparing VIPs with matched non-VIP controls, we will test whether VIPs are enriched for positive selection, either weak or strong, within specific epochs of human history.

These analyses will reveal whether viral adaptation in humans was episodic or continuous, and whether it was driven primarily by strong or weak selection. I expect to confirm the enrichment of adaptation in VIPs while extending prior work by identifying time-specific bursts of selection and polygenic signals. Beyond VIPs, this framework can be applied to other gene categories (e.g., immune, metabolic) and to trait-associated variants from GWAS, offering a generalizable pipeline for dissecting the modes and timing of adaptation across the genome.

***The role of regulatory elements in response to viruses***

Human virus-interacting proteins (VIPs) are central to adaptation against viruses, but the role of their regulatory elements (REs), such as those controlling gene expression, remains unresolved. To address this, I will investigate natural selection in putative REs of VIPs compared to non-VIPs. If REs are functionally important for responding to viruses, I expect REs of VIPs to show stronger conservation across species and stronger recent negative selection in humans. Alternatively, REs of VIPs may show limited cross-species conservation yet remain under strong negative selection in humans, suggesting that optimal regulatory strategies shift frequently while regulation itself remains critical. Conversely, no difference between VIPs and non-VIPs would suggest that protein-coding changes, rather than regulatory changes, dominate human adaptation to viruses.

In my previous work2, I curated ~4,000 VIPs and matched non-VIPs with equivalent levels of factors that influence selection on proteins. Building on this, I will classify REs, including enhancers, promoters, and other transcribed regions, using both physical proximity to genes and ChromHMM functional annotations12. To quantify constraint in REs, I will apply conservation scores such as PhastCons and PhyloP. I have also developed pipelines to infer the full distribution of fitness effects in human noncoding regions, which I will leverage in this analysis.

These results will advance our understanding of the evolutionary mechanisms underlying host responses to viruses and, more broadly, reveal how the immune system operates and how both regulatory elements and proteins adapt to strong environmental pressures.

**Future direction 2: Natural selection in humans and its connection to chronic diseases**

***DFE of non-coding regions across different human populations***

A central challenge in the era of precision medicine is understanding how consistent mutational effects are across human populations. While many studies show that a mutation’s impact can depend on genetic background, we still lack a systematic picture of these differences. One promising approach is to compare the distribution of fitness effects (DFE) across populations. Prior work 13 suggests that nonsynonymous mutations have broadly similar DFEs between species, but such analyses are limited to proteins, which are constrained by physical and chemical properties. In contrast, regulatory and other non-coding elements may evolve more flexibly. Moreover, studies of proteins are averaged across the entire genome, masking the possibility that some genomic regions experience local adaptation or evolve rapidly under environment-specific selection pressures.

Using genomic data from the Yoruba population, we have already inferred the DFE of putatively functional non-coding elements. We now aim to extend this analysis to other populations, such as Europeans, where most disease studies are concentrated. By jointly inferring demographic history and fitness effects, we will test how natural selection correlates across populations and identify how selection on regulatory elements in key biological pathways differ among populations. Ultimately, this research will illuminate how population-specific evolutionary histories shape the genetic architecture of chronic diseases, offering new insights into why certain populations may be more vulnerable to chronic diseases such as diabetes, cardiovascular disease, and autoimmune disorders.

***Phenotypic interpretation of selection in conserved and non-conserved noncoding genomic regions***

Genetic studies have revealed transitions of natural selection14, 15, yet the human phenotypes underlying these shifts remain unclear. My prior work showed that negative selection in human populations is stronger in noncoding regions conserved among primates but not across all mammals, compared to regions conserved across mammals but not in primates, and that substantial negative selection also operates in non-conserved regions.

To identify which human phenotypes contribute to these transitions, I will compare heritability enrichment and the distribution of fitness effects of trait-associated variants across regions with different levels of evolutionary conservation. Using GWAS summary statistics (UK Biobank and others), population variation data, and tools such as LDSC, SBayesS, and ASSESS, I aim to clarify how the genetic basis of traits relates to evolutionary constraint and to reveal chronic disease phenotypes shaped by lineage-specific selection in humans.

***Publicly accessible, variant-centered database of selection statistics***

Large databases such as dbSNP and gnomAD provide extensive human polymorphism data but lack integrated evolutionary information. In addition, existing resources with selection statistics do not leverage modern methods such as ML and ARG-based approaches and are increasingly outdated19, 20. Despite their advances, ML and ARG methods remain computationally demanding and underutilized in anthropology and biomedicine. I will generate and integrate selection statistics—including allele ages, selection coefficients, and timing of selection—for human SNVs across diverse populations using up-to-date methods. This resource will democratize access to evolutionary inference, foster cross-disciplinary connections with biomedical research, and link evolutionary dynamics to external annotations such as protein stability, mutational scans, and disease associations.