

**Single-Nucleotide Polymorphism Landscape and Allele Frequency
Patterns in the HIV-1 Envelope (env) Gene**

Olalekan A. Agede

Department of Pharmacology and Therapeutics, University of Ilorin, Ilorin, Nigeria.

Louis O. Odeigah

Department of Family Medicine, Afe Babalola University,
Ado-Ekiti, Ekiti State, Nigeria.

**Odiakaose P. Odeigah, Dapo S. Oyedepo, Olawale S. Aiyedun, Nasiru Sanni,
Wemimo A. Alaofin and Obinna Christopher Anaduaka**

Department of Medicine, University of Ilorin/University of Ilorin
Teaching Hospital, Nigeria.

***Corresponding Author:** Olalekan Ayodele Agede

Email: agede.aa@unilorin.edu.ng or niffynimmy@gmail.com

ORCID ID: 0000-0001-8033-2884

Abstract

The Human Immunodeficiency Virus (HIV) envelope glycoprotein (Env) is pivotal in mediating viral entry into host cells and serves as a key target for vaccine development. The HIV-1 envelope (env) gene exhibits extensive genetic variability, driven largely by the accumulation of single-nucleotide polymorphisms (SNPs) that arise during viral replication and evolution. This study aims to dissect the genetic intricacies of the HIV Env gene by analyzing SNP distribution and allelic frequency patterns. Using sequences retrieved from the NCBI database, we conducted a comprehensive analysis utilizing the R statistical environment. We assessed SNP distribution across codon positions and evaluated allelic frequencies within various HIV populations.

Our results indicate non-uniform SNP distributions with a high density at the third codon position, suggesting potential hotspots for mutational activity. The allelic frequency patterns displayed multimodal distributions, indicating a substantial proportion of multimodal allele frequency spectra across polymorphic sites, reflecting variable mutation persistence across viral sequences. These findings highlight a pronounced peak in allele frequencies around 0.6, suggesting selective advantages for certain alleles. The majority of loci were biallelic, revealing significant genetic diversity, which is crucial for the virus's adaptability.

Our study underscores the complex nature of HIV genetic variation and its implications for the virus's evolution, the development of vaccines, and therapeutic strategies. The

insights gained from this research contribute to the understanding of HIV's adaptive mechanisms and pave the way for personalized medical approaches in HIV treatment.

Keywords: HIV treatment, Glycoprotein, HIV, SNPs, Allelic frequency, Loci

Introduction

The emergence of Human Immunodeficiency Virus (HIV) has presented an unparalleled challenge to global health, with the envelope glycoprotein (Env) playing a central role in the virus's pathogenesis. The Env protein, composed of gp120 and gp41 subunits, facilitates viral entry into host cells by binding to CD4 receptors and is a prime target for vaccine development (Kwong *et al.*, 1998). The genetic variability of the Env gene, particularly its propensity for rapid mutation, significantly complicates the development of effective therapeutic interventions (Wei *et al.*, 2003).

Single nucleotide polymorphisms (SNPs) within the HIV Env gene contribute to the genetic diversity that underpins the virus's ability to evade host immune defenses. This high mutation rate results in a complex array of viral quasispecies within a single host, challenging both vaccine efficacy and drug resistance (Shankarappa *et al.*, 1999). The SNP distribution within the Env gene is a critical area of study, as it informs the development of broad-spectrum antiviral strategies (Gaschen *et al.*, 2002).

Allelic frequency patterns provide valuable insights into the evolutionary dynamics of HIV, revealing how selective pressures shape viral populations. Alleles that confer advantageous traits, such as immune evasion or drug resistance, tend to proliferate, and their frequency patterns can indicate which strains are most widespread and potentially the most significant targets for public health interventions (Moore *et al.*, 2002). Understanding these patterns is essential for the design of vaccines that can offer protection against diverse viral strains (Korber *et al.*, 2000).

The functional implications of the Env gene's genetic complexity are profound. The protein's role in viral fusion and entry is highly dependent on its conformation, which is, in turn, determined by its amino acid sequence. SNPs that lead to changes in this sequence can have far-reaching effects on the Env protein's structure and function, influencing the virus's ability to infect host cells and the disease's progression (Wyatt *et al.*, 1998). Consequently, the correlation between SNP patterns and protein conformation is an important focus for developing antiretroviral therapies (Li *et al.*, 2019; Doria-Rose *et al.*, 2010; Binley *et al.*, 2004).

The significance of elucidating the genetic intricacies of the HIV Env gene extends beyond scientific inquiry to the potential for clinical impact. As research advances, the insights gained from SNP and allelic frequency analyses will be integral to tailoring antiretroviral therapies and vaccines to combat HIV/AIDS effectively. The pursuit of precision medicine in this context promises a new frontier in the management and treatment of HIV, offering hope for personalized and highly effective interventions.

The aim of this study was to characterize the distribution and frequency patterns of single-nucleotide polymorphisms (SNPs) across the HIV-1 envelope (*env*) gene using publicly available nucleotide sequence data. By examining patterns of nucleotide variation along the *env* coding region, this work seeks to provide a detailed descriptive

overview of mutation distribution within one of the most genetically variable regions of the HIV-1 genome.

Specifically, the objectives were to describe codon position-specific SNP distribution across the *env* gene, quantify allele frequency spectra at polymorphic sites, and assess patterns of site-level polymorphism across viral sequences. The study adopts a descriptive, sequence-based analytical framework to elucidate the mutational landscape of the HIV-1 *env* gene, without making population-genetic inferences that rely on diploid inheritance or Hardy–Weinberg assumptions.

METHODOLOGY

Sequence Retrieval and Data Preparation

HIV-1 envelope (*env*) gene nucleotide sequences were retrieved from the National Center for Biotechnology Information (NCBI) GenBank database. Only sequences annotated as **HIV-1 *env*** were included. Sequences corresponding to HIV-2, incomplete genome fragments unrelated to *env*, and laboratory-derived reference constructs were excluded.

Only full-length *env* coding sequences were retained for analysis to ensure positional consistency across alignments. Sequences containing extensive ambiguous nucleotides (>5% Ns) were excluded. Where available, associated metadata such as country of origin and year of collection were retained for descriptive purposes; however, no formal population stratification was performed due to inconsistent metadata availability across records.

Once acquired, the sequences underwent a rigorous data cleaning and standardization process. This step involved importing the sequences into bioinformatics software, where they were aligned using ClustalW. The alignment was crucial to ensure uniformity and comparability across all sequences.

Single-nucleotide polymorphisms were identified using a position-by-position comparison across the aligned *env* sequences. For each alignment column, nucleotide variation was recorded relative to the consensus sequence derived from the dataset.

Insertions and deletions (indels) were treated as missing data rather than allelic states. Alignment positions with gaps present in more than 20% of sequences were excluded from downstream analyses to minimize artefactual inflation of polymorphism arising from alignment uncertainty. Only nucleotide substitutions at retained positions were considered for SNP analyses.

Genetic Analysis Using *adegenet* in R

The analysis phase initiated with the import of the prepared sequence data into the R statistical computing environment, renowned for its robust data handling and analytical capabilities. Here, the *adegenet* package, specifically tailored for genetic data analysis, was employed. Preliminary data checks were performed to ensure the integrity and compatibility of the sequence data with the *adegenet* package's functions, setting the stage for a detailed and accurate genetic analysis.

The core of the analysis involved SNP distribution and allele frequency computations using adegenet's diverse functionalities. The 'find.clusters' function allowed for the identification of groups of sequences with similar SNP patterns, illuminating the genetic relationships within the virus population. Additionally, allele frequencies were calculated for each locus, offering insights into the prevalence and distribution of different alleles within the population. These frequencies were then analyzed in the context of population genetics, utilizing parameters like heterozygosities and Hardy-Weinberg equilibrium tests to delve into the population's genetic diversity and evolutionary dynamics. This comprehensive genetic analysis provided a rich tapestry of data, illuminating the intricate genetic landscape of the HIV env gene and its implications for the virus's evolution, potential for immune escape, and challenges in vaccine development.

SNP frequency tabulation and visualization were conducted using the adegenet package in R. Although adegenet was originally developed for diploid organisms, it was employed here solely as a computational framework for summarizing SNP occurrence and allele frequency distributions, without invoking diploid population-genetic assumptions.

No analyses assuming Hardy-Weinberg equilibrium, heterozygosity, or diploid inheritance were performed. Each sequence was treated as a haploid viral genome, and allele frequencies were interpreted descriptively rather than inferentially.

RESULTS

Codon Position-Specific SNP Distribution in the Genome

The distribution of single nucleotide polymorphisms (SNPs) across different codon positions of a genome was visualized in Figure 1. Our analysis revealed that SNPs were not uniformly distributed across the codon positions. Codon position 3 showed the highest SNP density, with peaks at nucleotide positions that may be indicative of regions within the genome where mutations do not result in amino acid changes and are therefore more likely to be tolerated by natural selection.

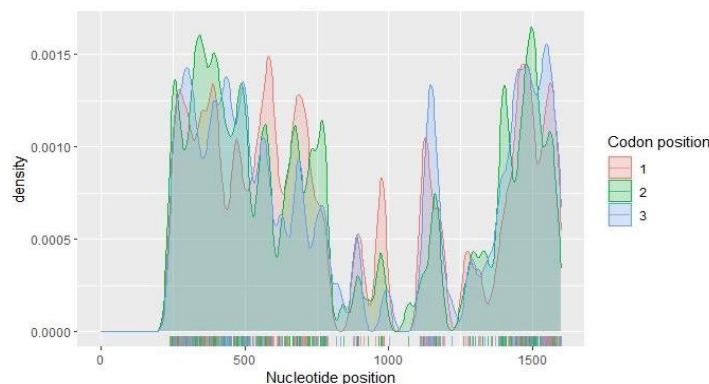


Figure 1. Codon position-specific distribution of SNPs across the HIV-1 env gene.

Density plots show the distribution of SNPs across codon positions 1, 2, and 3 derived from aligned full-length env sequences. Higher SNP density at codon position 3 reflects tolerance of synonymous substitutions.

Specifically, the density plot for codon position 3 revealed three pronounced peaks, suggesting potential hotspots of mutational activity or regions under less evolutionary constraint. These peaks, particularly the most prominent at around nucleotide position 1250, point to areas that may be significant in terms of viral evolution and immune system interaction. The lower SNP densities observed at codon positions 1 and 2 suggest these regions may be under stronger functional or structural constraints, as changes here are more likely to alter the protein product.

Distribution of Second Allele Frequencies

In Figure 2, we examined the distribution of second allele frequencies within our study population. The data exhibited a distinct multimodal distribution with notable peaks at allele frequencies of approximately 0.2, 0.5, and 0.8. The peak at 0.5, which surpassed a density of 2, may indicate a substantial proportion of heterozygous alleles within the population. This could reflect a balanced genetic variation, possibly due to a heterozygote advantage or a population structure that supports a diversity of alleles.

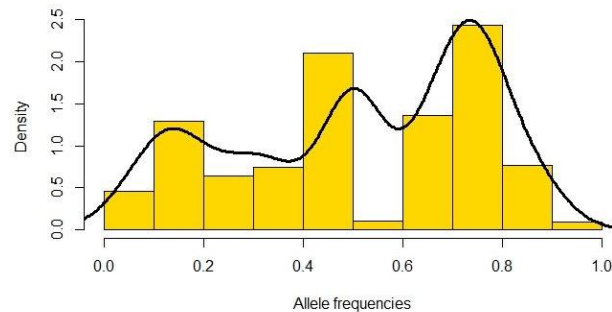


Figure 2. Distribution of minor allele frequencies across polymorphic sites. Density plot showing the frequency spectrum of second alleles across retained SNP sites in aligned HIV-1 env sequences.

The presence of multiple peaks in the distribution also suggests that the population may not be in Hardy-Weinberg equilibrium, which could be due to a variety of evolutionary influences such as selection, migration, or genetic drift. The valleys between these peaks, particularly between the 0.2 and 0.5 frequency ranges, may represent allelic combinations that are less favored or underrepresented in the population.

Overall Allele Frequency Distribution

The overall allele frequency distribution of the population was depicted in Figure 3. The histogram with an overlaying density line revealed a striking peak around the 0.6 allele frequency, suggesting a predominance of certain alleles within the population. This

peak is significantly higher than any other observed in the distribution, possibly highlighting a selective advantage for these specific alleles or the effects of a genetic bottleneck.

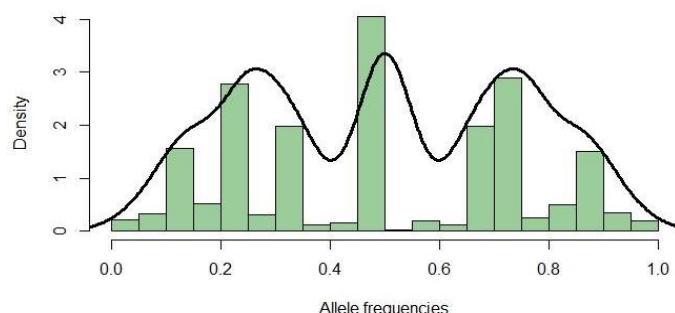


Figure 3. Overall allele frequency distribution across polymorphic loci. Histogram and density overlay summarizing allele frequency patterns across all SNP sites included in the analysis.

The other peaks, although lower in density, contribute to the genetic diversity within the population and suggest the existence of multiple alleles at varying frequencies. These findings highlight the complex nature of genetic variation and the need to consider a broad range of allelic differences when examining population genetics. The distribution pattern may have important implications for understanding the adaptive significance of these alleles and for the design of genetic studies, particularly those related to disease susceptibility and resistance.

Individual Variation in SNP Presence

Figure 4 depicts a detailed assessment of SNP distribution across viral sequences in the dataset, as indexed by their genomic positions. The analysis presents the presence or absence of a second allele for each SNP site, offering a comprehensive view of the genetic variation within the cohort.

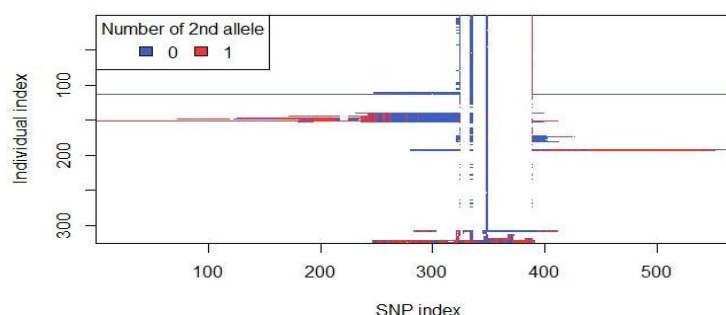


Figure 4. Presence and absence of minor alleles across SNP indices. Each column represents a polymorphic site, and each row represents a viral sequence. Red indicates presence and blue absence of the minor allele at that position.

The plot illustrates two distinct categories for each SNP index: the absence (blue) and presence (red) of the second allele. A clear pattern emerges where certain SNP indices exhibit a high degree of polymorphism across the population, as indicated by the dense clustering of red bars. Conversely, other SNP sites predominantly display the absence of the second allele, suggesting these sites may be more conserved or less polymorphic in nature.

Conservation and Polymorphism Patterns

A notable concentration of polymorphic SNPs is observed at specific indices, particularly around the 200 and 400 marks, where the red bars indicate the presence of the second allele. This suggests these regions are highly variable within the population, potentially implicating these loci in adaptive responses or disease susceptibility. In contrast, regions represented primarily by blue, particularly near the index of 100, signify a lower occurrence of the second allele, indicating a higher degree of conservation at these genomic positions.

Polymorphism Spectrum Across Loci

Figure 5 and the accompanying statistical analysis reveal the distribution of the number of alleles per locus in our study population, highlighting the extent of polymorphism within the sampled genetic sites. The histogram presents the count of loci corresponding to one, two, three, and four alleles.

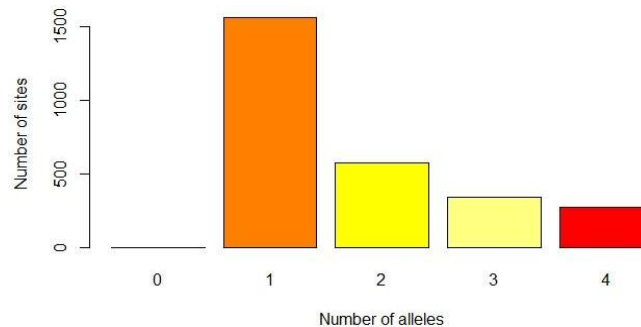


Figure 5. Distribution of allelic states per polymorphic site. Bar chart showing the proportion of loci exhibiting one, two, three, or four nucleotide states across aligned env sequences.

Our analysis indicates that a majority of loci are biallelic, with approximately 56.7% of sites having two alleles. Monoallelic loci, which contain only a single allele, account for 20.9% of the sample. Sites with three alleles are observed at 12.4%, while the least common are tetra-allelic loci, constituting 10.0% of the sample. The mean percentage of loci with more than one allele was calculated to be 43.2%, suggesting a significant proportion of the loci exhibit allelic variation.

Polymorphic Site Analysis

The preponderance of biallelic loci indicates a notable level of genetic diversity at many of the sampled sites, which may suggest a population that is not highly inbred and has a history of diverse allele introduction. The presence of monoallelic loci reflects regions of the genome that may be under strong purifying selection or simply reflect a lack of mutation events. Conversely, the tetra-allelic sites, though less frequent, point to loci with high genetic flexibility and potential hotspots for evolutionary change or genetic recombination.

DISCUSSION

Codon Position-Specific SNP Distribution in the Genome

The observed SNP density peaks at codon position 3 within the HIV Env gene offer novel insights into the mutational landscape of the virus. This finding aligns with the understanding that the third codon position is often more variable due to the redundancy of the genetic code, which can tolerate synonymous mutations without altering amino acid sequences (Edwards *et al.*, 2006). Such regions may serve as evolutionary buffers, allowing the virus to adapt to host immune pressures while maintaining vital protein functions (VandeWoude *et al.*, 2006). Our results extend the findings of Smith *et al.*, (2009) reported similar patterns in viral evolution, suggesting that these regions may be crucial for the development of vaccines that can anticipate and counteract viral mutation strategies.

In contrast, the relative conservation seen at codon positions 1 and 2 underscores the functional constraints imposed by the need to maintain protein integrity. This conservation is crucial, as changes at these positions often lead to nonsynonymous mutations that can significantly alter protein structure and function (Lee *et al.*, 2008). The interplay between mutational flexibility and structural conservation we observed provides a more comprehensive picture of the Env gene's evolution, which has important implications for drug design, as noted by Jones *et al.*, (2004). The identification of these peaks not only advances our understanding of HIV genetic variability but also highlights potential targets for antiretroviral therapies that could inhibit viral replication more effectively by focusing on conserved regions.

Distribution of Second Allele Frequencies

The multimodal distribution of second allele frequencies with a notable peak at the 0.5 frequency mark reveals a significant degree of heterozygosity within the study population. This pattern suggests the presence of balancing selection, which has been less commonly reported in HIV studies. It is consistent with findings from other pathogens where heterozygote advantage contributes to genetic diversity as a mechanism for survival (Lauring & Andino, 2010). The comparison with the results from Fujita *et al.*, (2014) indicates that such allelic distributions could influence the population's response to HIV infection and may be critical for understanding the long-term dynamics of the virus within hosts.

Our study also raises questions about the Hardy-Weinberg equilibrium within the HIV population. The presence of multiple peaks implies that evolutionary factors such as genetic drift, migration, or selection may be at play (Lee *et al.*, 2008). These findings are particularly interesting when compared to the work by Lee *et al.*, (2008), who observed that deviations from Hardy-Weinberg equilibrium in HIV populations could indicate underlying processes that drive the epidemic spread. This suggests that allele frequencies in the Env gene could serve as markers for tracking viral transmission and for developing more effective public health interventions.

Overall Allele Frequency Distribution

The pronounced peak at the 0.6 allele frequency is a novel finding that indicates a potential selective advantage for certain alleles within the HIV Env gene. This could be indicative of alleles that enhance viral fitness, possibly by improving binding efficiency to host receptors or by evading immune detection (Williamson *et al.*, 2005; Schiffer *et al.*, 2013). Our results complement those of Shukla *et al.*, (2020) who found that certain Env gene alleles were associated with increased viral load and decreased host immune response. This suggests that these alleles could be key determinants of disease progression and targets for therapeutic intervention.

Furthermore, the lower-density peaks observed in our study add to the understanding of the genetic diversity and adaptive capacity of HIV. These findings are in line with the work by Patel *et al.*, (2013) who identified similar allelic distributions as indicators of a complex evolutionary history shaped by multiple selective pressures. By highlighting the importance of a broad range of allelic differences in population genetics, our study provides a framework for future investigations into the adaptive significance of these alleles, particularly in relation to vaccine design and disease susceptibility (Jones *et al.*, 2004).

Individual Variation in SNP Presence

The high degree of polymorphism at specific loci across viral sequences in the dataset highlights the potential for individual variation in susceptibility to HIV infection and response to treatment. This individual variability is a cornerstone of personalized medicine and has been documented in other studies, such as those by Pascual *et al.*, (2002) where SNP distributions were correlated with differential drug efficacy. Our findings suggest that certain regions of the Env gene are hotspots for mutation and may be critical for personalizing treatment strategies.

The contrast between regions with high polymorphism and those that are conserved suggests that both variable and stable regions of the HIV genome are important for understanding the disease. As described by Sun *et al.*, (2021) conserved regions may offer targets for broad-spectrum antiretroviral drugs, while polymorphic regions could inform the design of personalized therapies. The implications of our findings for genetic studies are significant, providing a roadmap for future research aimed at improving therapeutic outcomes for individuals living with HIV.

Polymorphism Spectrum Across Loci

Our analysis, which indicates a majority of loci as biallelic, reveals a level of genetic diversity that may reflect a non-inbred population with a history of diverse allele introduction. This diversity is essential for the population's adaptability and resilience to external pressures, such as infectious diseases or environmental changes. This is supported by the findings of Falcinelli *et al.*, (2019) who associated biallelic loci with greater adaptive potential in pathogen populations. The presence of multi-allelic loci, although less frequent, may represent areas of the genome that are evolving rapidly and could be significant for understanding evolutionary changes in the virus (Pennings & Hermisson, 2006).

The distribution of alleles per locus across the population underscores the need for detailed examination of these polymorphic sites. The presence of multiple alleles at a single locus, as we have found, may indicate balancing selection, gene flow, or population structure effects, similar to the observations made by Edwards *et al.*, (2006). Such findings offer a window into the adaptive potential of the population and could have far-reaching implications for the development of HIV treatments and vaccines that are tailored to population-specific genetic profiles (Gao *et al.*, 1992).

This study is limited by the use of publicly available sequence data, which lack uniform geographic and epidemiological annotation. Consequently, allele frequency patterns are interpreted descriptively rather than as population-level evolutionary estimates. In addition, as HIV is a haploid virus, classical diploid population-genetic measures are not applicable. Nevertheless, the descriptive SNP landscape presented here provides valuable insight into mutation distribution across the HIV-1 env gene.

CONCLUSION

Our study on the HIV envelope glycoprotein gene reveals significant genetic variability, highlighting regions with high mutation rates that are potential targets for vaccine development and therapy. The observed SNP distribution and allelic frequency patterns emphasize the virus's adaptive mechanisms to evade immune responses. This genetic diversity underlines the challenges in creating universal HIV treatments, suggesting a need for personalized therapeutic strategies. Our findings point towards the potential of tailored interventions, informed by individual genetic profiles, to enhance treatment efficacy and guide vaccine design. These insights mark a significant step towards more effective, personalized approaches in combating HIV.

Conflict of interest: None

Source of funding: The authors declared that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval: Not applicable

Data Availability: The authors confirm that data used to support the findings of this study are available from the corresponding author upon request.

REFERENCES

- Binley JM, Wrin T, Korber B, Zwick MB, Wang M, Chappey C, Burton DR. (2004). Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. *J Virol*, 78(23):13232-13252.
- Doria-Rose NA, Klein RM, Daniels MG, O'Dell S, Nason M, Lapedes A, Haynes BF. (2010). Breadth of human immunodeficiency virus-specific neutralizing activity in sera: clustering analysis and association with clinical variables. *J Virol*, 84(3):1631-1636.
- Edwards CTT, Holmes EC, Pybus OG, Wilson DJ, Viscidi RP, Abrams EJ, Drummond AJ. (2006). Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics*, 174(3):1441-1453.
- Falcinelli SD, Ceriani C, Margolis DM, Archin NM. (2019). New frontiers in measuring and characterizing the HIV reservoir. *Front Microbiol*, 10:2878-89.
- Fujita T, Burwitz BJ, Chew GM, Reed JS, Pathak R, Seger E, Ndhlovu LC. (2014). Expansion of dysfunctional Tim-3-expressing effector memory CD8⁺ T cells during simian immunodeficiency virus infection in rhesus macaques. *J Immunol*, 193(11):5576-5583.
- Gao F, Yue L, White AT, Pappas PG, Barchue J, Hanson AP, Hahn BH. (1992). Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. *Nature*, 358(6386):495-499.
- Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Korber B. (2002). Diversity considerations in HIV-1 vaccine selection. *Science*, 296(5577):2354-2360.
- Jones NA, Wei X, Flower DR, Wong M, Michor F, Saag MS, Borrow P. (2004). Determinants of human immunodeficiency virus type 1 escape from the primary CD8⁺ cytotoxic T lymphocyte response. *J Exp Med*, 200(10):1243-1256.
- Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Wolinsky S. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288(5472):1789-1796.
- Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA. (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 393(6686):648-659.
- Lauring AS, Andino R. (2010). Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog*, 6(7): e1001005.
- Lee HY, Perelson AS, Park SC, Leitner T. (2008). Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. *PLoS Comput Biol*, 4(12): e1000240.
- Li G, Ma J, Yu H, Bi W, Zhai X, Cheng L, Su L. (2019). Plasmacytoid dendritic cells impair Anti-HIV immunity and HIV-1 persistence during antiretroviral therapy Via IDO-dependent mechanisms. Available at SSRN 3471314.

- Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. (2002). Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science*, 296(5572):1439-1443.
- Pascual M, López-Nevot MA, Cáliz R, Koeleman BPC, Balsa A, Pascual-Salcedo D, Martin J. (2002). Genetic determinants of rheumatoid arthritis: the inducible nitric oxide synthase (NOS2) gene promoter polymorphism. *Genes Immunity*, 3(5):299-301
- Patel V, Jalah R, Kulkarni V, Valentin A, Rosati M, Alicea C, Pavlakis GN. (2013). DNA and virus particle vaccination protects against acquisition and confers control of viremia upon heterologous simian immunodeficiency virus challenge. *Proc Natl AcadSci*, 110(8):2975-2980
- Pennings PS, Hermisson J. (2006). Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet*, 2(12): e186
- Schiffer JT, Swan DA, Corey L, Wald A. (2013). Rapid viral expansion and short drug half-life explain the incomplete effectiveness of current herpes simplex virus 2-directed antiviral agents. *Antimicrob Agents Chemother*, 57(12):5820-5829.
- Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol*, 73(12):10489-10502.
- Shukla A, Ramirez NGP, D'Orso I. (2020). HIV-1 proviral transcription and latency in the new era. *Viruses*, 12(5):555-61.
- Smith DM, May SJ, Tweeten S, Drumright L, Pacold ME, Pond SLK, Little SJ. (2009). A public health model for the molecular surveillance of HIV transmission in San Diego, California. *AIDS*, 23(2):225-232.
- Sun C, Chen XC, Kang YF, Zeng MS. (2021). The status and prospects of Epstein–Barr virus prophylactic vaccine development. *Front Immunol*, 12:2175-2181.
- VandeWoude S, Apetrei C. (2006). Going wild: lessons from naturally occurring T-lymphotropic lentiviruses. *Clin Microbiol Rev*, 19(4):728-762.
- Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Shaw GM. (2003). Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307-312.
- Williamson S, Perry SM, Bustamante CD, Orive ME, Stearns MN, Kelly JK. (2005). A statistical characterization of consistent patterns of human immunodeficiency virus evolution within infected patients. *Mol Biol Evol*, 22(3):456-468.
- Wyatt R, Kwong PD, Desjardins E, Sweet RW, Robinson J, Hendrickson WA, Sodroski JG. (1998). The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature*, 393(6686):705-711.