Single Nucleotide Polymorphism Induces a Positive Selection Pressure at Gag-Pol Sites in Human Immunodeficiency Virus Favoring Drug Resistance Mutations

Amreen Zahra1, Mushtaq A. Saleem1*, Hasnain Javed2, Muhammad Azmat Ullah Khan3, Muhammad Naveed1 and Abdul Rauf Shakoori4

1Department of Biochemistry and Biotechnology, University of Central Punjab, Lahore 54782, Pakistan
2Advanced Diagnostic Laboratory, Punjab AIDS Control Programme, Government of Punjab, Lahore 54000, Pakistan
3Department of Biochemistry and Molecular Biology, University of Gujrat, Gujrat 50700, Pakistan
4School of Biological Sciences, University of the Punjab, Quaid-I-Azam Campus, Lahore 54590, Pakistan

ABSTRACT

Human immunodeficiency virus the retrovirus causes AIDS. HIV mainly targets the body’s immune system, including the CD4 cells, which help the immune system fight off infections. Gag-Pol are the principal viral elements in HIV-1, which cause disease initiation and progression. Here we present a comprehensive study, where both the in vitro and in-silico analyses were performed on HIV-1 from isolates of the Punjab area of Pakistan. The main objective in the current study was to study the hypervirulence of SNPs in the Gag-Pol proteins by molecular modeling approaches. Mutational analysis in our study revealed S61A, S61M, S61Y, S61G, S61Q and M90L as the most hypervirulent mutations. This induces a selection pressure and a rate of increased virulence on Gag-Pol cleavage sites. These results significantly highlight the fact that the identified SNPs possibly contribute towards a positive selection pressure contributing to the identification of novel mutations like S61A, S61M and M90L at Gag-Pol cleavage sites harboring dominant drug resistance mutations. 3D structure validation demonstrated 87.3% in Gag and 92.5% in Pol in the most favored regions. Dolutegravir, a protease inhibitor, was used for molecular docking to reveal the drug resistance sites among the isolates under investigation. However, strikingly the binding energies resulted were -7.9 kcal/mol for Pol and -5.7 kcal/mol in Gag that predicted that SNPs in the Pol gene could be one of the reasons for drug resistance and the primary site of mutations progressing the onset of infection. These computational strategies will thus help in performing large scale studies and also in the development of novel medicines for the cure of HIV-1.

INTRODUCTION

Acquired immunodeficiency syndrome (AIDS) is initiated by the human immunodeficiency virus belonging to the Retroviridae family. Human immunodeficiency virus (HIV) mainly leads to reduced immunity, provoking increased vulnerability to opportunistic infections like HCV HBV, sexually transmitted diseases and also a variety of co-infections with HIV (Chang et al., 2013). HIV-1 may have originated in African from non-human primates and transferred to humans early in the 20th century (Worobey et al., 2008). HIV-1 and HIV-2 are the two main types that initiate the onset of infection. HIV-1 having its dwelling roots from Southern Cameroon was transferred to humans from wild chimpanzees during the twentieth century (Gao et al., 1999; Keele et al., 2006). Its evolution was from a simian immunodeficiency virus. On the other hand, HIV-2 seems to have its roots in Mangabey (Cercocebusatsys), Gabon, and Cameroon (Reeves and Doms, 2002).

The viral genome contains genes for the regulatory proteins such as Gag, Pol, and Env in addition to accessory and regulatory proteins (tat, rev, nef, vpr, vif, vpu). Gag encodes the matrix proteins p17, spacer protein 1 p2,
 spacer peptide 2, p1 nucleocapsid protein p7, capsid protein p24, p1 and P6 protein (Freed, 1998). While, Pol encodes RT, RNase H, integrase and protease. Cleavage of viral polyproteins, reverse transcription and integration into the host genome during the life cycle of HIV is done with the help of these viral enzymes (Mushahwar, 2006). Viral Gag-Pol precursor proteins are known to play an elemental role in the viral life cycle. These proteins thus facilitate the formation of mature virus particles through proteolytic processing (Fun et al., 2012). Functional variants can have both damaging and neutral effects on the structure and function of a protein. The hypervirulent nature of a protein provokes its destabilization, altering the gene regulation, thus posing a risk in maintenance of structural integrity of cells (Arshad et al., 2018). To address the question of these fatal anomalies, clinical trials and cohorts involved in Gag-Pol genotypic testing would be highly beneficial for therapeutic analysis. However, computational analysis of such genes would play a significant role in revealing underlying mechanisms and understanding patterns of drug resistance and mutations. Antiretroviral therapy (ART) is the only remedy ever since the diagnosis of this fatal virus. It favors by reducing high viral loads to undetectable levels, thus recovering the immune system. But, the increased virulency in mutations promotes drug resistance to ART which causes high rise in viral load rendering disruptions of the immune system.

Pakistan is working efficiently to fight and combat this deadly disease, which has increased in a couple of years. A sudden outbreak of HIV was reported in Kot Imrana, a small village located near Sargodha in Central Punjab in March 2018 (Zaid and Afzal, 2018). In 2019 more individuals were affected, indicating the spread of the disease to about 669 (13.38%) HIV patients and manifested increased mortality (Wahid, 2019). This crisis was followed by two major HIV outbreaks in 2019 in Rotoderu among infants and sudden increased cases of HIV-1 in Faisalabad, which was an alarming health catastrophe in Pakistan (Zahra et al., 2019; Mir et al., 2020).

To date, there is not much comprehensive molecular and computational analysis of HIV-1 in Pakistan, which could provide better insights into acquisition risks, transmission and challenges to HIV treatment. This is the first study in Pakistan on HIV-1, which covers a comprehensive in vitro and in-silico analysis targeting the Gag-Pol proteins in Punjab province of Pakistan. Since little is known about HIV-1’s molecular virology and bioinformatics, the present study was aimed at analyzing the two integral genes Gag-Pol of HIV-1 at molecular and computational level using various bioinformatics strategies to detect the increased virulence and deleterious effect of the mutations. Mechanisms underlying virulence and drug resistance need to be investigated. 3D models of the Gag-Pol protein will provide better insight for future studies in HIV and may prove helpful in development of medicines for the treatment of such non-treatable anomalies caused by genomic variations.

MATERIALS AND METHODS

Sample collection

This study was done in collaboration with Punjab AIDS Control Programme (PACP). Blood samples were obtained from different treatment centers of PACP, Lahore between November 2018 to October 2019. A total of 365 participants were recruited for this study, who had a viral load exceeding 100,000 copies/ml and a first CD4 cell count >350 cells/µl on the diagnosis. Antiretroviral (ART) naïve and ART therapy were included in the sample pool. A total number of 85 (23.29%) samples were confirmed with HIV-1 infection. Fifty-two males (14.25%) and 30 (9.04%) females were diagnosed with progressive HIV-1 infection. Only 33 HIV-1 (0.1%) positive isolates fulfilled the quality processing criteria for this study. Recently infected were 23 and 10 were already on ART treatment. The successfully amplified samples for HIV-1 Gag and Pol were used for genotyping and phylogenetic analysis. For detailed mutational and phylogenetic analysis, only 11 isolates qualified depending upon sociodemographic criteria mentioned in Table I. Sample collection was permitted and facilitated by PACP and informed consent was taken from the enrolled subjects. Study approval was granted by the Research Ethics Committee of the University of Central Punjab (Supplementary Fig. 2A, 2B).

Immunoassays and reference laboratory tests

All the samples were centrifuged for 3 min at 12,000 rpm followed by serum separation. The screening was then performed for the presence of HIV-1 infection by rapid testing devices. A three device method was used according to the guidelines suggested by WHO for diagnosis of HIV-1 in resource-limited countries (i) Alere™ HIV Combo Rapid Test (Alere Medical Co. Ltd., Matsudo-chi, Japan), (ii) Uni-Gold HIV (Uni-Gold, Trinity Biotech, Ireland; #1206502) and (iii) SD Bioline HIV-1/2 3.0 (Standard Diagnostics, Kyonggi-do, South Korea).

CD4 count testing and HIV detection by real-time PCR

After collection of the blood samples, 25 μl of blood was added to the Alere PIMA™ CD4 cartridge and CD4 count was displayed on the Alere PIMA™. Real-Time PCR was performed on the cobas 4800 HIV-1, COBAS®
AmpliPrep/COBAS® TaqMan® HIV-1 system, an in vitro nucleic acid amplification test for the quantification of HIV-1 RNA in plasma following automated specimen processing, amplification and detection. The test quantitates HIV-1 RNA at a range of 20-10,000,000 copies/mL (33 to 1.67x 10^7 International Units IU/mL) based on the WHO international standards recommended for HIV-1.

Table I. Baseline characteristics and sociodemographics of HIV-1 individuals.

| Gender                  |       |
|-------------------------|-------|
| Male                    | 7     |
| Female                  | 3     |
| Transgender             | 1     |

| Socioeconomic profile   |       |
|-------------------------|-------|
| Employed                | 5     |
| Unemployed              | 0     |
| Students                | 1     |
| Labourers               | 2     |
| People with very low income | 3 |

| Predicted risk factors for infection |       |
|-------------------------------------|-------|
| Alcohol intake                      | 0     |
| Sharing of needles and syringes      | 4     |
| Heterosexual contact                 | Unknown |
| Mother to child transmission         | 0     |
| Men who have sex with Men (MSM)      | Unknown |
| Truck drivers (travel history)       | 2     |
| Unknown                              | 5     |

| Urban/ rural population              |       |
|-------------------------------------|-------|
| People residing in cities            | 7     |
| People residing in village/ suburbs  | 4     |

| CD4 count (cells/μl) |       |
|----------------------|-------|
| < 200                | 6     |
| 200-350              | 5     |
| 350-500              | 0     |
| >500                 | 0     |

Nucleic acid extraction, reverse transcription, polymerase chain reaction and sequencing

Viral RNA was extracted from 200 μl of plasma by QIAamp Viral RNA Mini kit (Qiagen, Valencia, CA, United States) according to the manufacturer’s instructions. Revert Aid First Strand cDNA Synthesis Kit (Thermo Scientific, USA, # K1622) was used for cDNA preparation. Nested polymerase chain reaction (PCR) was used for the amplification of Gag and Pol genes. Partial Pol gene containing protease (PR) gene of 99 amino acid codons and the first 298 codon segment of the reverse transcriptase (RT) was amplified. Both Gag and Pol fragments were amplified by nested RT-PCR the method described by Chen et al. (2014). 1% agarose gel was used for visualization of the PCR products. Successfully amplified samples were sequenced by First Base Laboratories Sdn Bhd, Seri Kembangan, and Selangor, Malaysia on an ABI 3730xl automated DNA analyzer (Applied Biosystems, Foster City, CA).

Computational analysis

The sequences of Gag and Pol were submitted to Genbank (accession numbers MT 462197, MT 462198, MT 462199, MT 462200, MT 489306, MT 489307, MT 489308, MT 489303, MT 489304, MT 489305, MT 462202) in FASTA format using BankIt-NCBI-NIH database. Mutational analysis for Gag-Pol was performed by POLYPHEN 2, PROVEAN, SIFT and PHD-SNP to predict the allele function, effect of non-synonymous SNPs, structure and function-based characteristics (Adzhubei et al., 2013; Choi and Chan, 2015; Kamaraj and Purohit, 2013; Kumar et al., 2012). POLYPHEN 2 uses the naïve Bayesian classifier to predict the allele function with a score ranging from 0-1 calculating the probability for a given mutation to be deleterious mutation and a score nearer to 1 (Adzhubei et al., 2013). The prediction of conserved domains was performed by Pfam (Finn et al., 2015), CDD (Marchler-Bauer et al., 2014) and Scan Prosite (Hulo et al., 2006). To predict a protein sequence variation and its function, Protein Variation Effect Analyzer (PROVEAN) was used. When the score equalizes to a predefined threshold i.e. -2.5; the protein variant is found to have a more hypervirulent effect causing damage (Capriotti et al., 2006). The impact of non-synonymous SNPs on the structure and function of proteins was determined by using I-MUTANT 3.0. Sorting intolerance from tolerance (SIFT) is a web-based tool using the PSI-BLAST protein database. Functionally associated protein sequences were collected using SIFT. Due to the difference in software and algorithms, their predictions may differ from each other (Ng and Henikoff, 2001). PHD-SNP and SVM based classifier, classifies substitutions at specified positions, amino acid replacements and profile-based prediction. It classifies the mutations as neutral or deleterious polymorphism (Finn et al., 2015). Pfam database was used for analysis for the large collection of protein families represented by multiple sequence
alignments (MSA) and hidden Markov models (HMMs). To reveal the biomolecular interactions and mechanisms, docking methodologies are widely applied to structure-based drug design investigations. Further the visualization of the protein–ligand interactions were made by the PyMOL software. Computational docking for the study of biomolecular interactions and mechanisms was done by AutoDock Vina (AutoDock 4) (Bhargava et al., 2014). Phylogenetic tree analysis was performed using the neighbor-joining method with 1000 bootstrap replicates using MEGA (Molecular Evolutionary Genetics Analysis, version 5.1) (Tamura et al., 2011). RAMPAGE (Ho and Brasseur, 2005) and PROCHECK (Laskowski et al., 1993) tools were used for checking stereochemical quality, analyzing overall residue-by-residue geometry.

Table II. Mutational analysis using POLYPHEN-2, PROVEAN, I-MUTANT, PHD-SNP, SIFT.

| Gag gene | POLYPHEN-2 Mutations | PROVEAN Mutations | I-MUTANT (stability) | SIFT | PHD-SNP |
|----------|----------------------|-------------------|----------------------|------|---------|
| A219P    | Specific             | Neutral           | Decrease             | Tolerated | Diseased |
| 0.112    | 1.558 8              | 0.05              | 9                    |
| C263I    | Specific             | Deleterious       | Decrease             | Tolerated | Diseased |
| 0.224    | -4.282 5             | 0.10              | 2                    |
| N239K    | Specific             | Neutral           | Decrease             | Tolerated | Diseased |
| 0.161    | -3.456 2             | 0.91              | 2                    |
| A222G    | Specific             | Neutral           | Decrease             | Tolerated | Diseased |
| 0.977    | 2.321 10             | 0.34              | 1                    |
| S61A     | Sensitive            | Deleterious       | Decrease             | Tolerated | Diseased |
| 1.23     | -4.322 6             | 0.43              | 7                    |
| S61V     | Specific             | Deleterious       | Decrease             | Tolerated | Diseased |
| 0.076    | -1.256 6             | 0.89              | 7                    |
| S61L     | Not predicted        | Neutral           | Decrease             | Tolerated | Diseased |
| 0.00     | 1.549 8              | 0.54              | 3                    |
| S61I     | Specific             | Neutral           | Decrease             | Tolerated | Neutral |
| 0.433    | 1.985 3              | 0.65              | 2                    |
| S61M     | Sensitive            | Deleterious       | Decrease             | Tolerated | Diseased |
| 1.35     | -2.456 6             | 0.67              | 4                    |
| S61F     | Specific             | Neutral           | Decrease             | Tolerated | Diseased |
| 0.441    | 1.890 5              | 0.79              | 2                    |
| S61W     | Specific             | Neutral           | Decrease             | Tolerated | Neutral |
| 0.569    | 1.786 6              | 0.78              | 1                    |
| S61Y     | Sensitive            | Deleterious       | Decrease             | Tolerated | Diseased |
| 1.78     | -2.456 7             | 0.34              | 0                    |
| S61G     | Sensitive            | Deleterious       | Decrease             | Tolerated | Diseased |
| 1.27     | -4.578 9             | 0.54              | 5                    |
| S61P     | Specific             | Neutral           | Decrease             | Tolerated | Neutral |
| 0.344    | 0.433 4              | 0.73              | 0                    |
| S61S     | Specific             | Neutral           | Decrease             | Tolerated | Neutral |
| 0.442    | 2.654 9              | 0.87              | 5                    |
| S61T     | Specific             | Neutral           | Decrease             | Tolerated | Neutral |
| 0.554    | 1.983 8              | 0.94              | 5                    |

| Gag gene | POLYPHEN-2 Mutations | PROVEAN Mutations | I-MUTANT (stability) | SIFT | PHD-SNP |
|----------|----------------------|-------------------|----------------------|------|---------|
| S61C     | Specific             | Neutral           | Decrease             | Tolerated | Diseased |
| 0.221    | 1.976 5              | 0.58              | 2                    |
| S61H     | Specific             | Diseased          | Decrease             | Tolerated | Neutral |
| 0.119    | -0.987 7             | 0.49              | 0                    |
| S61R     | Specific             | Deleterious       | Decrease             | Tolerated | Diseased |
| 0.004    | -1.432 7             | 0.67              | 5                    |
| S61K     | Sensitive            | Deleterious       | Decrease             | Tolerated | Diseased |
| 1.44     | -1.987 9             | 0.87              | 0                    |
| S61Q     | Not predicted        | Neutral           | Decrease             | Tolerated | Neutral |
| 0.00     | 1.566 7              | 0.76              | 1                    |
| S61E     | Sensitive            | Deleterious       | Decrease             | Tolerated | Diseased |
| 1.28     | -2.334 6             | 0.65              | 1                    |
| S61N     | Sensitive            | Deleterious       | Decrease             | Tolerated | Diseased |
| 1.37     | -2.987 6             | 0.94              | 2                    |
| S61D     | Specific             | Neutral           | Decrease             | Tolerated | Neutral |
| 1.34     | -1.322 5             | 0.36              | 1                    |
| V81I     | Sensitive            | Deleterious       | Increase             | Not tolerated | 0.98 |
| 1.89     | -2.890 2             | 0.55              | 4                    |
| I116N    | Specific             | Neutral           | Decrease             | Tolerated | Diseased |
| 0.445    | 1.876 5              | 0.54              | 0                    |
| K115Q    | Specific             | Neutral           | Decrease             | Tolerated | Diseased |
| 0.655    | 2.345 3              | 0.34              | 3                    |
| M1141    | Sensitive            | Deleterious       | Increase             | Not tolerated | 0.93 |
| 1.34     | -1.445 3             | 0.95              | 4                    |
| D113E    | Specific             | Deleterious       | Increase             | Not tolerated | 0.99 |
| 1.99     | -1.322 7             | 0.94              | 4                    |
| G125S    | Specific             | Neutral           | Decrease             | Tolerated | Diseased |
| 0.886    | -2.12 9              | 0.26              | 2                    |
| T170S    | Specific             | Neutral           | Decrease             | Tolerated | Neutral |
| 0.645    | 1.987 8              | 0.54              | 4                    |
| N239T    | Not predicted        | Neutral           | Decrease             | Tolerated | Neutral |
| 0.00     | 0.223 2              | 0.67              | 3                    |
| G248S    | Sensitive            | Deleterious       | Decrease             | Not tolerated | 0.91 |
| 1.325    | -1.876 8             | 0.91              | 1                    |

| Pol gene | Mutations | I-MUTANT (stability) | SIFT | PHD-SNP |
|----------|-----------|----------------------|------|---------|
| M90L     | Sensitive | Deleterious          | Increase | Not tolerated | 0.96 |
| 1.88     | -1.997 3  | 0.96                | 2    |
| M143L    | Specific  | Deleterious          | Decrease | Tolerated | Neutral |
| 0.457    | -2.23 2   | 0.31                | 5    |
| L242I    | Specific  | Neutral              | Increase | Tolerated | Diseased |
| 0.644    | 1.739 0   | 0.49                | 6    |
| D269E    | Sensitive | Neutral              | Increase | Not tolerated | 0.95 |
| 1.47     | 1.677 5   | 0.76                | 3    |
| D277E    | Specific  | Neutral              | Decrease | Tolerated | Diseased |
| 0.876    | 1.889 7   | 0.76                | 5    |
| I278M    | Not predicted | Neutral | Decrease | Tolerated | Diseased |
| 0.00     | 0.999 5   | 0.87                | 0    |
| T100I    | Sensitive | Deleterious          | Increase | Not tolerated | 0.89 |
| 1.58     | -1.028 5  | 0.89                | 2    |
| D350E    | Specific  | Neutral              | Decrease | Tolerated | Neutral |
| 0.875    | 1.442 7   | 0.76                | 6    |
| Y221H    | Specific  | Neutral              | Decrease | Tolerated | Diseased |
| 0.710    | 1.133 4   | 0.65                | 4    |
| T160V    | Sensitive | Deleterious          | Decrease | Not tolerated | 0.98 |
| 1.20     | -2.559 6  | 0.98                | 5    |
RESULTS

Mutational analysis

We subjected all the detected single nucleotide polymorphisms (SNPs) of Gag and Pol to five different in silico approaches like POLYPHEN 2, I-MUTANT, SIFT, PROVEAN and PHD-SNP were used to investigate the effect on the structure of the Gag-Pol protein. However, four tools I-MUTANT, PHD-SNP, SIFT and PROVEAN strongly verified the prevailing hypervirulence in Gag and Pol with a higher degree of deleterious, specific and sensitive SNP values as shown in Table II. The identified SNPs demonstrated deleterious and damaging effects with the specified scores, thus depicting an increased influence of diseased SNPs shown in Table II.

POLYPHEN 2 (Polymorphism Phenotyping V2)

Almost the majority of the mutations in Polyphen 2 analysis showed a change in amino acid towards specificity and sensitivity, denoting that the mutation causes a deleterious and damaging effect with a score near to 1.00 and greater than 1 (Table II). The highest specificity was demonstrated at position S61 in Gag and M90L in Pol, giving a score of 1.88 (Table II). The mutational analysis of this research study revealed S61A, S61M, S61Y, S61G and S61Q were the most dysfunctional mutations in Gag and probably could be one of the reasons for increased drug resistance. The deleterious SNPs are based on position-specific independent count score difference, where score “1” is considered the most deleterious and our results also demonstrated synchroney as reported in some recent computational studies (Arshad et al., 2018).

PROVEAN

The cutoff for Gag variants C263I, S61A, S61N, V81I and in Pol T160V depicted a significant deleterious effect of more than -2.5 thus denoting as the most virulent mutations detected by Provean analysis (Table II).

I-MUTANT

In the current study, majority of the SNPs (Table II) in the Gag gene, demonstrated a decreased score of greater than “5” significantly denoting the fact that the disease could be caused by any of these mutations. SNPs D269E and Y221H showed a score of about 5 and 8, respectively, in the Pol gene exhibiting deleterious effect with increased virulent influence. The output predicted the decreased changes in protein stability upon the possible mutations for a given position with RI value (Reliability Index) 1 and 2 for the transitions, as mentioned above. Our results revealed that most of the SNPs detected showed decreased stability in I-mutant detected Gag-Pol variants (Table II).

In accordance with these results, previous studies have also demonstrated decreased protein stability leading to degradation, misfolding and aggregation of proteins (Mayer et al., 2007). This phenomenon of decreased stability, however, sheds light on the notion that these detected polymorphisms probably cause the maximum harmful effect to the protein by affecting its stability and disrupting the proper translation of the protein.

PHD-SNP

The output of Gag-Pol variants by PHD-SNP (Table II) contains mutations in the protein sequence. A219P, a Gag variant, demonstrated mutation with a score of 9 and L242I, a Pol detected variant gave a score of 6.

Conserved domains and motifs prediction

For the prediction of conserved domains, the following tools were used:

Pfam

The domains predicted by Pfam for Gag-Pol from the 11 isolates were Gag_p17, Gag_p24, zf-CCHC, RVT-1, RV-T thumb, RV-T connect, RNase-H, Integrase-Zn and MLVINC as shown in Figure 1. The matrix proteins of Gag showed an E-value of 1.6e-61 and 4.8e-79, whereas Pol demonstrated an E-value of 8 i.e. e-31and and 5.6e-18 of RV-T1 and RV-T thumb (Supplementary Table S1). The significant domains of Pol were protease, RNase H and Integrase, which denoted the E-values 1.1e-24, 3.3e-11 (Supplementary Table S1). These results infer the fact that the major mutations in both these genes could be in these detected domains with raised E-values suggesting a higher rate of virulence in mutations of these genes. Highly conserved residues could be functional or structural based on their location on the protein surface or inside its core (Doniger et al., 2008). Amino acids involved in vital processes, for example, in interactions among different proteins, are more conserved than others. SNPs located at the conserved regions are considered immensely deleterious to protein as compared to those at non-conserved sites (Miller and Kumar, 2001).

CDD

Results of conserved data domain (CDD) revealed 10 major conserved domains named as RT-rtv, Gag_p24, Gag_p17, RTV connect, RNase-H, RVP, rve, RV-T-thumb (Figs. 2A and 2B). The highest E-values were depicted by Pol region of domains RV-T1 and RV-T thumb with 8.1e-31 and 5.6e-18. These domains majorly contributed to the catalytic activity of proteins at the N-terminal. Gag_p24, mainly involved in the formation of the inner protein layer, depicted an E-value of 1.36e-64, whereas Gag_p17, the matrix
Fig. 1. Sequence matches and features of Gag-Pol-pol as predicted by Pfam: 10 domains Gag_p17, Gag_p24, zf-CCHC, RVT-1, RVT_connect, RNase-H, RVP, RVT-thumb, MLVINC, Integrase-Zn were best predicted by Pfam as shown in the figure.

Fig. 2. A and B, Predicted conserved domains and E-values of Gag-Pol by CDD database of NCBI: Highly conserved regions of Gag_p17, Gag_p24, RVT-1, RVT_thumb were predicted in HIV-1. The domains for Pol showed an active site, DNA binding site, dNTP binding site and NNRTI binding site.

Table III. Conserved domains as predicted by CDD: Description analysis for Gag_p24 and Gag_p17 describing as core nucleocapsid and icosahedral proteins for the mature immunodeficiency virus with E-value as 1.36e-64 and 1.17e-20. RT_Rtv and RVT_thumb (four helix domain) catalyzing the major part of the virus from RNA to DNA for integration into the host depicted the major E-values as 5.51e-123 and 1.66e-12 respectively. The Interval values were 86-256, 1-74 for the matrix Gag proteins and 29-245, 254-292 for Pol proteins.

| Domain   | Accession ID | Description                                                                 | Interval | E-value  |
|----------|--------------|-----------------------------------------------------------------------------|----------|----------|
| Gag_p24  | c102908      | Gag gene protein p24 (core nucleocapsid protein) p24 forms inner protein layer of the nucleocapsid. ELISA tests for p24 is the most commonly used method to demonstrate virus replication both in vivo and in vitro. | 86-256   | 1.36e-64 |
| Gag_p17  | c102892      | The matrix protein forms an icosahedral shell associated with the inner membrane of the mature immunodeficiency virus. | 1-74     | 1.17e-20 |
| RT_Rtv   | cd01645      | Reverse transcriptases catalyze the conversion of single-stranded RNA into double-stranded viral DNA for integration into host chromosomes. Proteins contain long terminal repeats (LTRs) and are multifunctional enzymes with RNA-directed DNA polymerase, DNA directed DNA polymerase, and ribonuclease hybrid (RNase H) activities. The viral genome enters the cytoplasm as part of a nucleoprotein complex, and the process of reverse transcription generates in the cytoplasm forming a linear DNA duplex via an intricate series of steps. | 29-245   | 5.51e-123|
| RVT_thumb| c106055      | This domain is known as the thumb domain. It is composed of a four helix bundle. | 254-292  | 1.66e-12 |

Aspartyl proteases and RT-Pol exhibited a confidence level of (-1) and increased high scores (Supplementary Table S2). These predictions reveal information about the biochemical function of the core HIV-1 proteins and have also been described previously in other studies (Naveed et al., 2014).

Scan prosite
The predictions resulted from Scan Prosite were functional sites, protein families by comparing amino acid patterns in the Gag-Pol protein domains like ferredoxin type iron sulphur binding regions, EGF-1, Anaphylatoxin-1, Aspartyl proteases and RT-Pol exhibited a confidence level of (-1) and increased high scores (Supplementary Table S2). These predictions reveal information about the biochemical function of the core HIV-1 proteins and have also been described previously in other studies (Naveed et al., 2014).

Secondary structure prediction
The translate tool EXPasy was used to predict the secondary structure of the protein. The DNA sequence
was translated by EXpasy, further, the translated sequence was modeled with I-TASSER for tertiary structure prediction.

**Tertiary structure prediction**

The first model having the best-predicted scores gave a C-score of -1.18, TM-score 0.57±0.15 and an estimated RMSD 8.6±4.5Å in the Gag 3D predicted model (Supplementary Fig. 1A). The Pol 3D model showed a C-score of -0.49, TM-score of 0.65±0.13 and an estimated RMSD 7.6±4.3Å (Supplementary Fig. 1B). Similar studies done previously on GDH gene and mental retardation genes like TUSC3 have used I-TASSER for tertiary structure prediction (Naveed et al., 2014, 2016). These findings significantly strike the notion which would be highly beneficial in the development and understanding of novel therapeutic strategies for diseases.

**Phylogenetic analysis**

MEGA 5 software was used to elucidate the phylogenetics and ancestral history in the current study. The (Neighbor-Joining method) was used for the construction of a phylogenetic tree. The interior branch method and a bootstrap value of 1000 were implemented to validate the ancestral relationships (Figs. 3A and 3B).

**3D Structure validation**

**Ramachandran plot**

The 3D structure validation was done by RAMPAGE (http://services.mbi.ucla.edu/SAVES/Ramachandran/) and PROCHECK (Ho and Brasseur, 2005) tools which provide a view of the conformation of a protein and verification of the 3D structure. Gag had 87.3% residues in the most favored regions, 11.3% in additional allowed regions, 1.0% in generously allowed regions, 0.4% in disallowed regions. The total number of residues was 821 (Fig. 4B). Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected.
to have over 90% in the most favored regions.

**Fig. 4.** Rampage analysis: The Ramachandran plots show the phi (ϕ)-psi (ψ) torsion angles for every residue of a protein. A, The final model had 87.3% residues in the most favored regions, 11.3% in additional allowed regions. Total number of residues were 821. B, The final model had 92.5% residues in the most favored regions, 6.9% in additional allowed regions. Total number of residues were 971. Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

**Auto dock vina**

Tools such as Pymol, discovery studio and Mgl tool were used in docking methodology (Trott and Olson, 2010). Here we checked the binding capabilities between the protein of HIV and the most recommended drugs for HIV-1 like Dolutegravir (DTG) ligand possibly for control of AIDS. Molecular docking imaging analysis (Figs. 5 and 6) exhibited interactions visualized between the protein and ligand. Structural shreds of evidence demonstrated the hydrogen bonds and binding interactions both in Figures 5 and 6, respectively. Drugs like DTG have been well known for their inhibition reported in previous and current studies. DTG mainly functions by binding the catalytic domain of the HIV Gag-Pol region, diminishing the activity of the viral enzymes by the assembly of the HIV genome ceasing into the host which blocks the access of the substrate rendering the entry to the catalytic site (Wainberg et al., 2016). For the current research analysis, DTG was used as a major docking drug with the obtained I-TASSER structures of the isolates to check the ligand-binding sites and increased virulency among them and comparable with the wild type structure of HIV.

**DISCUSSION**

Principally in the mechanism, HIV-1 protease inhibitors (PIs) compete with the Gag substrates for attachment with the active site of the protease (da Silva et al., 2019). However, even when the enzyme exhibits reduced activity, secondary mutations develop in the protease gene, which fastens the process of Gag substrate cleavage, thus initiating viral fitness (Dam et al., 2009). Alternative theoretical methodologies have been proposed in the literature previously suggesting that mutations in the Gag-Pol region actively participated in the resistance against protease inhibitors (PI) and that is resulting in the virus resistant to these PIs in the absence of any major mutations in the protease gene (Nijhuis et al., 2007; Verheyen et al., 2006). Docking results showed binding affinities of both Gag and Pol with protease inhibitors like Dolutegravir gave a binding affinity of -5.7 kcal/mol in Gag and -7.9 kcal/mol in Pol (Supplementary Table S3A and S3B). This suggests that DTG one of the most potential drug for control of HIV. However the findings of this study also gives coverage to the fact that drug resistance could be mainly in the Gag-
Pol region of HIV-1 thus suggesting that these two proteins create a positive selection pressure at Gag-Pol cleavage sites leading to progression of the disease. However, the binding affinity of DTG was lower in Pol, which suggested significantly increased virulence in the Pol region and enhanced vulnerability to mutations with an up-regulated hypervirulence. The binding sites were very few or only in a single place, as shown in Figure 5A with lysine residues and with lysine and valine residues (Fig. 6A). In addition to this, the results of docking of this study also shed light on the phenomenon that in future therapeutic implications with synthetic compounds of DTG can be introduced in a combination of other drugs as the major protease inhibitors for ART therapy for Pakistani population.

Fig. 6. Molecular docking image analysis of Pol. The interactions are visualized between the protein and ligand of Dolutegravir with Pol protein. A shows the solid surface of the protein with residue labels Lysine and valine of amino acids. B represents the cartoon structure with meshy surface depicting the binding interaction of different residues of receptor with ligand atoms as shown. In figure C the cartoon structure and binding yellow lines show the hydrogen bonds. D demonstrates the meshy surface with dots.

These in vitro findings confirm the molecular revelations for the major mutations and underlines the importance of Gag-Pol hypervirulence in resistance in HIV-1. As future research perspectives, this must be regarded as a primary significant phase for determining the mutational effects, positive selection pressure and interactions on viral resistance and fitness. Hence, this study can pave beneficial computational strategies for mutational analysis in the cure of HIV-1 by drug production.

CONCLUSION

Nevertheless, the underlying mechanisms still remain unclear, how HIV-1 achieves the regulation of fusion, suggesting more future studies are required to define the mechanism of this regulation. These in vitro findings confirm the molecular revelations for the Gag-Pol major mutations and underlines the importance of Gag-Pol mutations in resistance, very similar to the results obtained in this study. As future research perspectives, this must be regarded as a major significant phase for determining the mutational effects and interactions on viral resistance and fitness. Furthermore, such comprehensive investigations of immune diseases will help in the exploration of pathways for drug development and innovative methods for saving lives and curing HIV. Hence, this study can pave beneficial computational strategies for mutational analysis in the cure of HIV-1 by drug production.

ACKNOWLEDGMENTS

Special thanks to Dr. Munir Ahmad Malik Project Director, and Dr. Asim Altaf Punjab AIDS Control Programme (PACP) for providing assistance and facilities at PACP. We are also very grateful for the contributions of the laboratory staff at PACP for providing the opportunity to collect the samples and laboratory procedures as recommended by WHO.

Compliance with ethical standards

This research study was ethically approved by the Institutional Review Board of University of Central Punjab and permitted by the Punjab AIDS Control Programme also.

Funding

The research study was funded by University of Central Punjab, Lahore and Punjab AIDS Control Programme.

Supplementary material

There is supplementary material associated with this article. Access the material online at: https://dx.doi.org/10.17582/journal.pjz/20200525130507
Statement of conflict of interest
The authors have declared no conflict of interest.

REFERENCES

Adzhubei, I., Jordan, D.M. and Sunyaev, S.R., 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, 76: 7.20.1-7.20.41. https://doi.org/10.1002/0471142905.hg0720s76

Arshad M., Bhatti, A. and John, P., 2018. Identification and in silico analysis of functional SNPs of human TAGAP protein: A comprehensive study. *PLoS One*, 13: e0188143. https://doi.org/10.1371/journal.pone.0188143

Bhargava, K., Nath, R., Seth, P.K., Pant, K.K. and Arshad M., Bhatti, A. and John, P., 2018. Identification of single point protein mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. *PLoS Pathog.*, 5: e1000345. https://doi.org/10.1371/journal.ppat.1000345

Doniger, S.W., Kim, H.S., Swain, D., Corcueria, D., Williams, M., Yang, S.P. and Fay, J.C., 2008. A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet.*, 4: e1000183. https://doi.org/10.1371/journal.pgen.1000183

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A. and Salazar, G.A., 2016. The Pfam protein families database: Towards a more sustainable future. *Nucl. Acids Res.*, 44: D279-285. https://doi.org/10.1093/nar/gkv1344

Freed, E.O., 1998. HIV-1 gag proteins: diverse functions in the virus life cycle. *Virology*, 251: 1-15. https://doi.org/10.1006/viro.1998.9398

Fun, A., Wensing, A.M., Verheyen, J. and Nijhuis, M., 2012. Human immunodeficiency virus Gag and protease: Partners in resistance. *Retrovirology*, 9: 63. https://doi.org/10.1186/1742-4690-9-63

Gao, F., Bailes, E., Robertson, D.L., Chen, Y., Rondenburg, C.M., Michael, S.F., Cummins, L.B., Arthur, L.O., Peeters, M., Shaw, G.M. and Sharp, P.M., 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes. *Nature*, 397: 436-441. https://doi.org/10.1038/17130

Ho, B.K. and Brasseur, R., 2005. The Ramachandran plots of glycine and proline. *BMC Struct. Biol.*, 5: 1-11. http://www.biomedcentral.com/1472-6870/5/14

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M. and Sigrist, C.J., 2006. The prosite database. *Nucl. Acids Res.*, 34: D227-D230. https://doi.org/10.1093/nar/gkj063

Kamaraj, B. and Purushot, R., 2013. In silico screening and molecular dynamics simulation of disease-associated nsSNP in TYRPI1 gene and its structural consequences in OCA3. *Biomed. Res. Int.*, 2013: Article ID697051. https://doi.org/10.1155/2013/697051

Keele, B.F., Van Heuserswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M.L., Bibollet-Ruche, F., Chen, Y., Wain, L.V., Liegeois, F., Loul, S., Ngole, E.M., Bienvenue, Y., Delaporte, E., Brookfield, J.F.,
Positive Selection Pressure at Gag-Pol Sites in Human Immunodeficiency Virus

Sharp, P.M., Shaw, G.M., Peeters, M. and Hahn, B.H., 2006. Chimpanzee reservoirs of pandemic and non pandemic HIV-1. Science, 313: 523–526.

Kumar, A. and Purohit, R., 2012. Computational screening and molecular dynamics simulation of disease associated nsSNPs in CENP-E. Mutat. Res., 738: 28-37. https://doi.org/10.1016/j.mrmm.2012.08.005

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M., 1993. PROCHECK: A program to check the stereochemical quality of protein structures. J. appl. Cryst., 26: 283-291. https://doi.org/10.1107/S002188989200944

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., Lanzczyk, C.J., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C. and Bryant, S.H., 2015. CDD: NCBI’s conserved domain database. Nucl. Acids Res., D1(2015): D222-D226. https://doi.org/10.1093/nar/gku1221

Mayer, S., Rüdiger, S., Ang, H.C., Joerger, A.C. and Fersht, A.R., 2007. Correlation of levels of folded recombinant p53 in Escherichia coli with thermodynamic stability of folded recombinant p53 and Fersht, A.R., 2007. Correlation of levels of folded recombinant p53 in Escherichia coli with thermodynamic stability of folded recombinant p53.

Miller, M.P and Kumar, S., 2001. Understanding human disease mutations through the use of interspecific genetic variation. Hum. mol. Genet., 10: 2319–2328. https://doi.org/10.1038/19435

Mir, F., Mahmood, F., Siddiqui, A.R., Baqi, S., Abidi, S.H., Kazi, A.M., Nathwani, A.A., Ladhani, A., Qamar, F.N., Soofi, S.B. and Memon, S.A., 2020. Covid-19: NCBI’s conserved domain database. Mol. Biol. Evol., 28: 2731–2739. https://doi.org/10.1093/molbev/msu121

Mushahwar, I.K., 2006. Human immunodeficiency viruses: Molecular virology, pathogenesis, diagnosis and treatment. Perspect. med. Virol., 13: 75-87. https://doi.org/10.1016/S0168-7069(06)13005-0

Naveed, M., Haider, I., Khalid, N. and Muntaz, A.S., 2014. Bioinformatics based structural characterization of glucose dehydrogenase (gdh) gene and growth promoting activity of Leclercia sp. QAU-66. Braz. J. Microbiol., 45: 603-611. https://doi.org/10.1590/1517-83822014000200031

Naveed, M., Kazmi, S., Anwar, F., Arshad, F., Dar, T. and Zafar, M., 2016. Computational analysis and polymorphism study of tumor suppressor candidate Gene-3 for non syndromic autosomal recessive mental retardation. J. appl. Bioinformat. Comput. Biol., 5: https://doi.org/10.4172/2329-9533.1000127

Ng, P.C. and Henikoff, S., 2001. Predicting deleterious amino acid substitutions. Genome Res., 11: 863–874. https://doi.org/10.1101/gr.176601

Nijhuis, M., van Maarseveen, N.M., Lastere, S., Schipper, P., Coakley, E., Glass, B., Rovenska, M., de Jong, D., Chappey, C., Goedegebure, I.W., Heilek-Snyder, G., Dulude, D., Cammack, N., Brakier-Gingras, L., Konvalinka, J., Parkin, N., Kräusslich, H.G., Brun-Vezinet, F. and Boucher, C.A., 2007. A novel substrate-based HIV-1 protease inhibitor drug resistance mechanism. PLoS Med., 4: e36. https://doi.org/10.1371/journal.pmed.0040036

Reeves, J.D., and Doms, R.W., 2002. Human immunodeficiency virus type 2. J. Gen. Virol., 83: 1253-1265. https://doi.org/10.1099/0022-1317-83-6-1253

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S., 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony methods. Mol. Biol. Evol., 28: 2731–2739. https://doi.org/10.1093/molbev/msr121

Trott, O. and Olson, A.J., 2010. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J. Comput. Chem., 31: 455-461. https://doi.org/10.1002/jcc.21334

Verheyen, J., Litau, E., Sing, T., Däumer, M., Balduin, M., Oette, M., Fätenheuer, G., Rockstroh, J.K., Schuldenzucker, U., Hoffmann, D., Pfister, H. and Kaiser, R., 2006. Compensatory mutations at the HIV cleavage sites p7/p1 and p1/p6-gag in therapy-naive and therapy-experienced patients. Antivir. Ther., 11: 879–887.

Wahid, B., 2019. An update on the severe outbreak of HIV in Kot Imrana, Pakistan. Lancet Infect. Dis., 19: 241. https://doi.org/10.1016/S1473-3099(19)30038-6

Wainberg, M.A., Han, Y.S. and Mesplède, T., 2016. Might dolutegravir be part of a functional cure for HIV? Can. J. Microbiol., 62: 375-382. https://doi.org/10.1139/cjm-2015-0725

Worobey, M., Gemmel, M., Teuwen, D.E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.J., Kabongo, J.M., Kalengayi, R.M., Van Marck, E., Gilbert, M.T. and Wolinsky, S.M., 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature, 455: 661–664. https://doi.org/10.1038/nature07303

Might dolutegravir be part of a functional cure for HIV? Can. J. Microbiol., 62: 375-382. https://doi.org/10.1139/cjm-2015-0725

Worobey, M., Gemmel, M., Teuwen, D.E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.J., Kabongo, J.M., Kalengayi, R.M., Van Marck, E., Gilbert, M.T. and Wolinsky, S.M., 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature, 455: 661–664. https://doi.org/10.1038/nature07303

Might dolutegravir be part of a functional cure for HIV? Can. J. Microbiol., 62: 375-382. https://doi.org/10.1139/cjm-2015-0725
Zahra, A., Naveed, M., Ahmad, I., Saleem, M.A. and Zeshan, B., 2019. The syringes catastrophe of the HIV outbreak in Faisalabad, Punjab, Pakistan. *Future Virol.*, 2019: 571-572. https://doi.org/10.2217/fvl-2019-0084

Zaid, M. and Afzal, M.S., 2018. HIV outbreak in Pakistan. *Lancet Infect. Dis.*, 18: 601. https://doi.org/10.1016/S1473-3099(18)30281-0
Supplementary Material

Single Nucleotide Polymorphism Induces a Positive Selection Pressure at Gag-Pol Sites in Human Immunodeficiency Virus Favoring Drug Resistance Mutations

Amreen Zahra¹, Mushtaq A. Saleem¹*, Hasnain Javed², Mohammad Azmat Ullah Khan³, Muhammad Naveed¹ and Abdul Rauf Shakoori⁴

¹Department of Biochemistry and Biotechnology, University of Central Punjab, Lahore 54782, Pakistan
²Advanced Diagnostic Laboratory, Punjab AIDS Control Programme, Government of Punjab, Lahore 54000, Pakistan
³Department of Biochemistry and Molecular Biology, University of Gujrat, Gujrat 50700, Pakistan
⁴School of Biological Sciences, University of the Punjab, Quaid-I-Azam Campus, Lahore 54590, Pakistan

Supplementary Fig S1 A: I-TASSER protein predicted models, A: 3D view of Gag protein, B: 3D view of Pol protein

* Corresponding author: dr.mushtaq@ucp.edu.pk

Copyright 2021 Zoological Society of Pakistan
The Human Research Ethics Committee  
Faculty of Life Sciences  
University of Central Punjab  
Lahore  

Subject: REQUEST FOR THE APPROVAL OF BLOOD SAMPLE COLLECTION FOR PHD PROJECT ON MOLECULAR AND COMPUTATIONAL ANALYSIS OF HIV-1 IN PAKISTAN  

Respected Sir/Madam,

It is stated that Miss Amreen Zahra is a PhD Research scholar (Biochemistry) at Faculty of Life Sciences, University of Central Punjab under my Supervision. Her PhD research topic is MOLECULAR AND COMPUTATIONAL ANALYSIS OF HIV-1 IN PAKISTAN. She has successfully defended the PhD synopsis and thus will do sample collection of HIV-1 with research biotechnologist expert under complete biosafety conditions. Kindly review the project for approval in the above mentioned topic of research and your approval is highly appreciated and I express my deep gratitude and thanks for it.

Thanking you very much in anticipation.

Your’s sincerely,

Supervisor:

Dr. Muhammad Naveed  
Assistant Professor  
FLS, UCP  

Approval by the Ethics Committee:

1- Member 01 Dr. Muhammad Saqib Shahzad (UCP-FLS)  
2- Member 02 Dr. Sumayya Nizamuddin (SKMCHRC)  
3- Member 03 Dr. Muhammad Faheem (AIMC)  
4- Member 04 Mr. Nadeem Alvi (UCP-Pharmacy)  
5- Member 05 Mr. Muhammad Shafeeq ur Rahman (UCP-Pharmacy)  
6- Member 06 Dr. Muhammad Naveed (UCP-FLS)
Supplementary Fig. S2. Approval of blood sample collection.
Supplementary Table S1. Conserved domains prediction by Pfam: The domains predicted by Pfam of Gag-Pol region, are Gag_p17, Gag_p24, zf-CCHC RVT-1, RVT_connect, RNase-H, RVP, RVT-thumb, MLVINC and Integrase-Zn.

| Source       | Domain     | Start | End  | Gathering threshold (bits) | Score (bits) | E-value      |
|--------------|------------|-------|------|---------------------------|--------------|--------------|
| Pfam         | Gag_P17    | 2     | 132  | 20.80                     | 10.00        | 218.50       |
|              |            |       |      |                           | 217.40       | 7.6e-62      |
|              | coiled_coil| n/a   | 87   | 121                       | n/a          | n/a          |
|              | disorder   | n/a   | 98   | 131                       | n/a          | n/a          |
|              | low_complexity | n/a | 104  | 119                       | n/a          | n/a          |
| Pfam         | Gag_P24    | 149   | 363  | 29.00                     | 29.00        | 277.70       |
|              |            |       |      |                           | 276.10       | 1.6e-79      |
|              | disorder   | n/a   | 175  | 176                       | n/a          | n/a          |
|              | disorder   | n/a   | 183  | 192                       | n/a          | n/a          |
|              | disorder   | n/a   | 197  | 200                       | n/a          | n/a          |
|              | disorder   | n/a   | 202  | 203                       | n/a          | n/a          |
|              | disorder   | n/a   | 211  | 249                       | n/a          | n/a          |
|              | disorder   | n/a   | 361  | 362                       | n/a          | n/a          |
|              | disorder   | n/a   | 374  | 378                       | n/a          | n/a          |
| Pfam         | zf-CCHC    | 390   | 407  | 20.80                     | 16.70        | 37.90        |
|              |            |       |      |                           | 25.10        | 2.9e-06      |
|              | coiled_coil| n/a   | 424  | 444                       | n/a          | n/a          |
|              | disorder   | n/a   | 447  | 481                       | n/a          | n/a          |
| Pfam         | RVT 1      | 650   | 821  | 29.60                     | 29.60        | 119.60       |
|              |            |       |      |                           | 118.40       | 3.4e-31      |
|              | Pfam       | RVT thumb | 828 | 891                       | 20.70        | 92.80        |
|              |            |       |      |                           | 90.80        | 1.7e-18      |
|              | low_complexity | n/a      | 883 | 891                       | n/a          | n/a          |
| Pfam         | RVT connect| 905   | 1006 | 21.00                     | 21.00        | 197.00       |
|              |            |       |      |                           | 196.10       | 1.1e-55      |
|              | low_complexity | n/a | 982  | 990                       | n/a          | n/a          |
| Pfam         | RNase H    | 1023  | 1144 | 25.80                     | 25.80        | 100.10       |
|              |            |       |      |                           | 98.20        | 3.1e-25      |
|              | Pfam       | Integrase Zn | 1155 | 1192                      | 21.70        | 55.70        |
|              |            |       |      |                           | 53.80        | 8.6e-12      |
|              | disorder   | n/a   | 1294 | 1296                      | n/a          | n/a          |
|              | disorder   | n/a   | 1298 | 1310                      | n/a          | n/a          |
|              | disorder   | n/a   | 1340 | 1341                      | n/a          | n/a          |
| Pfam         | Rve        | 1443  | 1558 | 32.50                     | 32.50        | 92.40        |
|              |            |       |      |                           | 90.30        | 5.5e-23      |
|              | disorder   | n/a   | 1549 | 1551                      | n/a          | n/a          |
|              | disorder   | n/a   | 1590 | 1591                      | n/a          | n/a          |
|              | disorder   | n/a   | 1607 | 1611                      | n/a          | n/a          |
|              | disorder   | n/a   | 1617 | 1618                      | n/a          | n/a          |
|              | low_complexity | n/a | 1619 | 1631                      | n/a          | n/a          |
| Pfam         | MLVINC C   | 1652  | 1730 | 29.20                     | 29.20        | 116.20       |
|              |            |       |      |                           | 115.00       | 1.3e-30      |
|              | disorder   | n/a   | 1709 | 1719                      | n/a          | n/a          |
|              | disorder   | n/a   | 1726 | 1731                      | n/a          | n/a          |
### Supplementary Table S2. Scanprosite results for Gag-Pol with domain descriptions.

| Domain                                      | Description                                      | Confidence level | Range | Sequence                        | Score   |
|----------------------------------------------|--------------------------------------------------|------------------|-------|---------------------------------|---------|
| 2FE2S_FER_12Fe-2S (ferredoxin-type iron-sulfur binding region signature) | EGF_1 (EGF-like domain signature 1)              | PS00197 (-1)     | 24-32 | CTGATCATC                      |         |
| ANAPHYLATOXIN_1 (Anaphylatoxin domain signature) | ANAPHYLATOXIN_1 (Anaphylatoxin domain signature) | PS01177 (-1)     | 530-562 | CCtaGgtgacat.GCTctgatg...TacCat- tgc.CC |         |
| ASP_PROT_RETROV (Aspartyl protease)          | ASP_PROT_RETROV (Aspartyl protease)              | PS50878 (-1)     | 16 - 85 | KEALLDTGADDTVLEDINLP-...MIGGIGGFIKVKQY-DHLIEICGKKAIGTVLGYTPVNYIQRNM |         |
| ANAPHYLATOXIN_1 (Anaphylatoxin domain signature) | RT_POL (Reverse transcriptase (RT) catalytic domain profile) | PS50878 (-1)     | 139 - 329 | EGKISRGIEPNYPYAIFIAKKK-KSTLWKEVDVRELNKRTQDFW-EVQLGHPAGLKKKSVTVLD-VGDAYESVPLYEDFRKYTAFT-IPstnepGVRYQYNVLQGWK-GSPAIFQSSMTKIPEFKQNYQndivl-YQYMDDLYVGSDLIEIGqHRTKIEELRGLHKLKWFGTPDDKHKQ | 32.439 |
| ASP_PROTEASE (Eukaryotic and viral aspartyl proteases active site) | ASP_PROTEASE (Eukaryotic and viral aspartyl proteases active site) | PS00141 (-1)     | 18 - 29 | ALLDTGADDTVLP                  |         |

### Supplementary Table S3. Molecular docking was done to predict bound conformations and free energies of binding of small-molecule ligands to macromolecular targets. A: Shows the binding energies of Gag with Dolutegravir and the binding energies predicted were -5.7 kcal/mol with distance from rmsd lower binding energy as 0.00 and best mode rmsd upper binding energies as 0.00. B: Shows the binding energies of Pol with Dolutegravir and the binding energies predicted were -7.9 kcal/mol with distance from rmsd lower binding energy as 0.00 and best mode rmsd upper binding energies as 0.00.

| Mode | affinity (kcal/mol) | distance from rmsd l.b. | best mode rmsd u.b. |
|------|---------------------|-------------------------|---------------------|
| 1    | -5.7                | 0.000                   | 0.000               |
| 2    | -5.7                | 14.813                  | 18.458              |
| 3    | -5.7                | 13.564                  | 17.660              |
| 4    | -5.7                | 19.777                  | 24.082              |
| 5    | -5.6                | 19.700                  | 24.163              |
| 6    | -5.6                | 20.369                  | 24.682              |
| 7    | -5.4                | 13.887                  | 17.966              |
| 8    | -5.4                | 18.641                  | 20.808              |

### Supplementary Table S3B. Binding affinity energies with Pol and Dolutegravir.

| Mode | affinity (kcal/mol) | distance from rmsd l.b. | best mode rmsd u.b. |
|------|---------------------|-------------------------|---------------------|
| 1    | -5.7                | 0.000                   | 0.000               |
| 2    | -5.7                | 14.813                  | 18.458              |
| 3    | -5.7                | 13.564                  | 17.660              |
| 4    | -5.7                | 19.777                  | 24.082              |
| 5    | -5.6                | 19.700                  | 24.163              |
| 6    | -5.6                | 20.369                  | 24.682              |
| 7    | -5.4                | 13.887                  | 17.966              |
| 8    | -5.4                | 18.641                  | 20.808              |
| 9    | -5.4                | 18.641                  | 20.808              |