Spontaneous Mutations in HIV-1 Gag, protease, RT p66 in the first replication cycle and how they appear: Insights from an *in vitro* BSL2 assay on mutation rates and types.

Joshua Yi Yeo[a], Darius Wen-Shuo Koh[a], Ping Yap[a], Ghin-Ray Goh[a],& Samuel Ken-En Gan[a,b,c,#]

[a]*Antibody & Product Development Lab, BII, Agency for Science, Technology and Research (A*STAR), Singapore 138672.*

[b]*Experimental Drug Development Centre, Agency for Science, Technology and Research (A*STAR), Singapore 138670*

[c]*p53 Laboratory, Agency for Science, Technology and Research (A*STAR), Singapore 138648.*

Running Title: HIV-1 Gag, protease, RT p66: Mutation rates and types

#Corresponding author: Samuel_Gan@eddc.a-star.edu.sg; samuelg@bii.a-star.edu.sg

Antibody & Product Development Lab, BII, A*STAR

60 Biopolis Street, #B2 Genome

Singapore 138672

Tel: +65 64070584

Keywords: HIV, mutation rate, Gag, protease, reverse transcriptase, drug resistance

Abstract

While drug resistant mutations in HIV-1 is largely credited to its error prone HIV-1 RT, host proteins such as deaminases may also play a role generating mutations. Many HIV-1 RT mutational *in vitro* studies utilize reporter genes (LacZ) as template, leaving out the possible contribution of HIV codon usage and gene-specific effects. To address this gap, we studied HIV-1 RT mutation rates and bias on its own Gag, protease, and RT p66 genes in an *in-vitro* selection pressure free system. We found rare clinical mutations with a general avoidance of crucial functional sites in the background mutations rates for Gag, protease and RT p66 at 4.71 x 10^{-5}, 6.03 x 10^{-5}, and 7.09 x 10^{-5} mutations/bp respectively. Gag and p66 genes showed a large number of 'A to G' hypermutations likely due to cellular adenosine deaminases. Comparisons with silently mutated p66 sequences showed an increase in mutation rates (1.88 x 10^{-4} mutations/bp) and that 'A to G' mutations occurred in regions reminiscent of ADAR neighbour preferences. Mutational free energies by the 'A to G' mutations revealed an avoidance of destabilizing effects with the natural p66 gene codon usage providing barriers to ADAR effects. Our study demonstrates the importance of studying mutation emergence in HIV genes to understand how fast drug resistance can emerge, sometimes with contributions of host deaminases, providing transferable applications to how new viral diseases and drug resistances can emerge.

Introduction

RNA viruses have been shown to have a higher likelihood of undergoing genetic changes, leading to species jump [1] and efficient spread among humans [2]. Notably, among RNA viruses, the human immunodeficiency viruses such as HIV-1 and HIV-2 are believed to be zoonotic transmissions of the simian immunodeficiency viruses (SIV) [3,4].

Within HIV-1, the Gag and protease proteins play crucial roles in viral assembly and maturation of infectious virions [5]. Protease cleaves Gag and Pol polyproteins into functional subunits [6] and this cleaving is prevented by protease inhibitors (PI) which compete with Gag for the active site [7,8]. In emerging PI resistance, mutations on viral protease reduce affinity to PIs [8,9], and the gradual accumulation of many such resistance mutations [10-12] induced by HIV-1 RT [13,14], eventually render the PIs ineffective, and in some cases, cross-resistances to limit clinical drug selection [8]. In many cases, mutations on the substrate Gag compensate for reduced viral fitness [8,9,14-16] with paired mutations that work synergistically to protease mutations [17-21] to overcome PIs.

In the spotlight for the cause of the mutations in HIV is the error prone enzyme reverse transcriptase (RT), an asymmetric heterodimer of the p66 and p51 subunits [22]. The p66 subunit performs the key enzymatic functions of catalysing DNA polymerisation and cleaving the RNA of the RNA/DNA duplex [23,24], while the p51 subunit plays a more supportive role to p66 [25].

In the investigation of HIV mutations, in-depth analysis of RT is indispensable. However, most of the previous studies on HIV mutation rates to date utilize reporter genes such as LacZ for the template for analysis of mutations, and not HIV genes. There is thus a current gap in understanding the contribution of HIV gene-specific sequences and codon usage in mutational hotspots, as well as type of mutations and when they can emerge in the infection cycle [26]. To fill this gap, we developed an *in-vitro* based assay without translational, immune

and drug selection pressures to characterize the basal HIV-1 RT mutations and biases on HIV-1 Gag, protease and RT p66 genes. As a control and further investigation to the natural codon usage, we also created a silent codon mutated variant of RT p66. Together, our findings shed light on the emergence of drug resistance mutations, the native rate, where they appear, and the associated biases and trends that would be useful for the design of future drug interventions.

Methods

Transfection of HIV-1 Gag, Pr and RT plasmids

HIV-1 Gag and Pr genes were PCR amplified from plasmid p8.91 [27]. HIV-1 RT p66 (GenBank: K03455.1) and codon mutated p66 genes were synthesized (BioBasic). The codon mutated p66 sequence was generated by reverse translation from the amino acid sequence for differing nucleotide (but retaining the amino acid sequence) to disrupt RNA secondary structures. The genes were then cloned separately into the pTT5 plasmid vector (YouBio), transformed into competent DH5α *E. coli* cells [28], and transfected into HEK293F cells as previously performed [29-31].

RNA extraction and cDNA synthesis

Total RNA from cells transfected two days prior with the gene of interest were extracted using TRIzol according to manufacturer's instructions (Invitrogen). cDNA synthesis was performed using recombinant HIV-1 RT subunits: (i) p51 (0.2475 µg) and p66 (0.2125 µg) from Sino Biological Inc (catalogue: 40244-V07E and 40244-V07E1, respectively), (ii) DNase-treated RNA, (iii) 5X RT buffer (25 mM Dithiothreitol, 375 mM KCl, 15 mM $MgCl_2$, 250 mM Tris-HCl [pH 8.3]), (iv) Oligo$(dT)_{18}$ (Thermo Scientific), (v) dNTPs, and (vi) RiboLock RNase inhibitor (ThermoFisher Scientific), in a single cycle of 25°C for 18 min, 37°C for 1 hr and 85°C for 5 min. RT negative controls were prepared without the addition of HIV-1 RT p51 and p66 subunits.

Amplification of cDNA and TOPO cloning

PCR amplification was performed using high-fidelity Q5 Polymerase PCR (New England Biolabs) with the following primers sets: HIV-1 Gag - F (5 '– TAT TAG GAA TTC ATG GGT GCG AGA GCG– 3') and R (5' – CTG GTA AAG CTT CTA GTG GTG GTG GTG – 3'); protease - F (5' – GCG GCC GAA TTC ATG CCT CAA ATC AC – 3') and R (5' – TAT AAT AAG CTT CTA GTG GTG GTG GTG –3'); RT p66 - F (5' – ATG GCC TTG ACC TTT GCT TTA CTG – 3') and R (5'– CTT GTC GTC ATC GTC TTT GTA GTC – 3'); and codon mutated p66 - F (5'– GCG GTG ATG GAT GGA CCA AAA GTA AA – 3') and R (5'– CTG CGC CTA ATG ATG ATG ATG AT – 3') as per manufacturer's recommendations. PCR products were analysed by gel electrophoresis with GelApp [32] and purified using Gel extraction and PCR purification kits previously described [33]. Purified PCR products were cloned using Zero Blunt TOPO PCR cloning kit (Invitrogen) as per manufacturer's protocol and transformed into in-house competent DH5α cells as previously described [28]. Transformed DH5α cells were plated and grown overnight at 37°C on LB agar plates supplemented with kanamycin (50 µg/ml). Transformants were screened using GoTaq PCR (Promega) with universal M13F (-20) forward and reverse primers prior to sequencing.

Sequence analysis

Sequence assembly and alignment of HIV-1 sequences were performed using YAQAAT Webserver [34]. DNA2App [35] was used to analyse nucleotide and amino acid sequences. Mutations in the cDNA gene sequences were determined by multiple sequence alignments with characterised HIV-1 sequences from the Los Alamos sequence database [36]. To rule out sequencing artefacts, sequence chromatograms were analysed, with repeat re-sequencings performed for ambiguous peaks and detected mutations. The mutation rates were

calculated as mutations/bp where the total number of mutations and the total nucleotide bases of the respective HIV-1 genes were compared.

A 'Two Sample logo' was generated to determine the underlying sequence contexts of 'A to G' transitions generated in our in-vitro assay for comparisons against previously reported ADAR neighbour preferences [37] . A custom script was written to locate, pool and align 4 nucleotides upstream and downstream of all adenosine mutants. Analysis of natural adenosines were also analysed to generate an ersatz set of sequences as background control (see Figure 4a for schematic representation). The mutated sites were compared against the respective ersatz background sequences using the Two Sample logo software [38] with two sample t-test without Bonferroni correction to test for significantly enriched and depleted bases within 9 nucleotides.

In-silico assessment on protein thermostability using FoldX and Rosetta Cartesian_ddg

To study the mutations structurally in Gag, *in silico* mutagenesis was performed on previously modelled compact Gag structures [13] using PyMOL followed by minimization using GROMOS96 implementation in Swiss-PdbViewer. For p66, the 3T19 crystal structure was used.

Protein thermostability, free energy changes of the mutations $\Delta\Delta G_{mt}$ ($\Delta G_{wild\text{-}type}$ - ΔG_{mutant}) were studied using FoldX5 and Rosetta Cartesian_ddg (Version : 2017.52.58848) [39]. Model structures were first relaxed and minimized using either the FoldX RepairPDB protocol or with the cartesian-space refinement relax protocol and ref2015_cart score function prior to mutagenesis. Free energy modelling was performed with the Cartesian_ddg protocol (https://www.rosettacommons.org/docs/latest/cartesian-ddG). To calculate the mutational effects on protein thermostability with FoldX, BuildModel (using default settings) was used with the numberOfRuns parameter set to 10. For Rosetta procedures, the lowest scoring structure out of 1000 cartesian-space minimization iterations were assessed to be free of serious

clashes by the MolProbity server [40]. The Cartesian_ddg protocol and talaris_cart score functions were used to generate 15 models for each mutation and the average of the lowest 3 scoring mutations were averaged and multiplied by the scaling factor [39] to calculate the $\Delta\Delta G_{mt}$ in kcal/mols.

Results

Characterization of HIV-1 Gag mutant variants

In our study, we identified and classified 'A to G' mutations as hypermutations when multiple mutations were found on the same sequence variant given that mutations should statistically be evenly distributed.

801 HIV-1 Gag sequences were generated and calculated to show a mutation rate of 3.36×10^{-5} mutations/bp and at 4.71×10^{-5} mutations/bp when including 'A to G' hypermutations (see Table 1).

More transitions (n = 24 excluding 'A to G' hypermutations and n = 40 including) than transversions (n = 12) were found (Figure 1A). 'A to G' substitutions were found to be the most frequent (n = 20, 37.7%) when including 'A to G' hypermutations, but when excluded, 'T to C' substitutions were the most frequent (n = 9, 25.0%), followed by 'G to A' and 'G to T' (n = 6, 16.7%), 'C to A' and 'C to T' (n = 5, 13.9%), 'A to G' (n = 4, 11.1%), and 'T to G' (n = 1, 2.8%), with several substitutions types not observed (see Figure 1B and C).

Amino acid analysis showed missense mutations (n = 30, 55.6%) to occur at approximately 2 times the frequency of silent mutations (n = 15, 27.8%, see Figure 2) while nonsense mutations (n = 5, 9.3%) and frameshift mutations (n = 4, 7.5%) had lower occurrences (see Table S1 for full list of mutations).

Characterization of HIV-1 protease mutant variants

For HIV-1 protease, 640 sequences were generated with a calculated mutation rate of 6.03×10^{-5} mutations/bp (see Table 1). No 'A to G' hypermutations were found for protease.

As with Gag, there were more transition mutations (n = 6) than transversions (n = 2) as shown in Figure 1A. 'G to A' substitutions were found to be the most frequent (n = 3, 37.5%), followed by 'A to G' (n = 2, 25%), 'A to C', 'T to C' and 'C to A' (n = 1, 12.5%), with several substitutions not observed (see Figures 1B and C).

Amino acid analysis showed that missense mutations (n = 4, 36.4%) occurred at approximately 1.3 times the frequency of silent mutations (n = 3, 27.3%) and frameshift mutations (n = 3, 27.3%), whereas nonsense mutations (n = 1, 9.1%) were of lower occurrences (see Figure 2, Table S2 for full list).

Characterization of HIV-1 RT p66 mutant variants

From 571 HIV-1 RT p66 subunit sequences, we calculated a mutation rate of 4.27×10^{-5} mutations/bp when excluding 'A to G' hypermutations and 7.09×10^{-5} mutations/bp when including them (see Table 1).

As with Gag and protease, there were more transition mutations (n = 18 excluding 'A to G' hypermutations, 45 including 'A to G' hypermutations) than transversions (n = 11) as shown in Figure 1A. 'A to G' substitutions were found to be the most frequent (n = 29, 51.8%) only when including 'A to G' hypermutations. When excluded, 'T to C' and 'G to T' substitutions were instead the most frequent (n = 8, 27.6%), followed by 'G to A' (n = 6, 20.7%), 'C to A' (n = 4, 12.5%), and 'A to G' and 'C to T' (n = 2, 6.9%). There were several substitutions not observed (see Figure 1B and C).

Analysis after translation found similar trends for both Gag and protease where missense mutations (n = 31, 48.4%) occurred at 1.6 times the frequency of silent mutations (n

= 19, 29.7%) as shown in Figure 2. Nonsense mutations (n = 2, 3.1%) and frameshift mutations (n = 12, 18.7%) were of lower occurrences (Table S3).

Characterization of HIV-1 Codon Mutated RT p66 mutant variants

A total of 700 codon mutated p66 gave a calculated mutation rate of 6.53 x 10^{-5} mutations/bp when excluding 'A to G' hypermutations and 1.88 x 10^{-4} mutations/bp when including them (see Table 1).

In the same trend with Gag, protease, and wild-type p66, there were more transition mutations (n = 37 excluding 'A to G' hypermutations, 176 including 'A to G' hypermutations) than transversions (n = 29) in codon mutated p66 (Figure1A). Of the same trend with wild-type p66, 'A to G' substitutions were found to be the most frequent (n = 146, 71.2%) only when including 'A to G' hypermutations, and when excluded, 'G to T' substitutions were found to the most frequent (n = 17, 25.8%), followed by 'G to A' (n = 13, 19.7%) and 'T to C' (n = 13, 19.7%), 'A to G' (n = 7, 10.6%), 'C to A' (n = 8, 12.1%), 'C to T' (n = 4, 6.1%), 'A to T' (n = 2, 3.0%), and 'T to A' and 'T to G' substitutions (n = 1, 1.5%), with the absence of the following substitutions: 'A to C', 'C to G' and 'G to C' (see Figure. 1B and C).

Amino acid analysis showed agreement in trends with the other HIV-1 genes where missense mutations (n = 109, 56.2%) occurred approximately 1.6 times the frequency of silent mutations (n = 70, 36.1%, see Figure 2). Nonsense mutations (n = 7, 3.6%) and frameshift mutations (n = 8, 4.1%) were at lower occurrences (Table S4).

Mutation Rates of the Domains of HIV-1 Gag, protease and RT p66

The HIV-1 genes were further analysed by their respective domains and the mutation rates calculated separately (see Figure 3). Within Gag, the overall mutation rates throughout the domains were within a narrow range of 2.29 to 4.46 x 10^{-5} mutations/bp when excluding 'A to G' hypermutations, but increased to 8.16 x 10^{-5} mutations/bp when including them. 'A to G' hypermutations were found in the capsid (CA), nucleocapsid (NC), and p6 subunits, but not in the matrix (MA), p2 and p1 domains. Mutations were not observed in the MA/CA ^{129}SQNY/PIV^{135}, CA/p2 ^{360}ARVL/AEA^{366}, p2/NC ^{374}ATIM/IQK^{380}, P1/p6 ^{445}PGNF/LQS^{451} cleavage sites and the ^{277}YSPTSIL^{283} capsid linker. However, a couple of mutations (A431D and F433L) occured in the NC/p1 ^{429}RQAN/FLG^{435} cleavage site.

For HIV-1 protease, there were no 'A to G' hypermutations and no mutations in the active site ^{25}DTG^{27}. For RT p66, when excluding hypermutations, there was a range of 0 to 9.12 x 10^{-5} mutations/bp, which increased to 1.35 x 10^{-4} mutations/bp when including them. Despite the 2nd finger domain being of similar length as the 1st palm segment, only 'A to G' hypermutations were observed in the 2nd finger without other mutations. No mutations were observed in the RT p51 and RNase H domain (p51-RNH) cleavage site AETF^{440}↓Y^{441}VD or in the catalytic triad: D110, D185 and D186. For the codon mutated p66, the mutation rate was in the range of 2.98 x 10^{-5} – 1.06 x 10^{-4} mutations/bp when excluding hypermutations and 3.38 x 10^{-4} mutations/bp when including them. Contrary to the wild-type p66, there were mutations observed in the 2nd finger domain, suggesting that codon usage of the wild-type 2nd finger domain may provide some barrier to mutations that are not 'A-G' hypermutations. Similarly, no mutations were observed in the p51-RNH cleavage site AETF^{440} ↓ Y^{441}VD despite the different codon usage. Interestingly, 'A to G' hypermutations were not observed in the 1st palm region of the codon mutated p66.

Sequence contexts of ‘A to G’ mutations

To determine the influence of host defence deaminases, specifically ADAR in the overrepresentation of ‘A to G’ mutations in our assay as was performed on double stranded non-HIV RNA [37], we analysed the neighbouring sequences +/-4 nucleotides (nt) of all the ‘A to G’ mutations using the two-sample logo analysis (Figure 4A-C).

Analysis of the Gag showed significant depletions of ‘A’ at -4 and ‘G’ at -1 nt positions downstream of ‘A to G’ mutation sites at ($p < 0.05$) agreeing with known ADAR neighbour preferences [37] (see Figure 4B). Using the Two Sample Logo of multiple and single substitution mutations, there was an enrichment of ‘C’ at the +1, +4 ($p < 0.05$) and -4 nt ($p < 0.01$) positions (see Figure S1), with the former two (+1 and +4 nt positions) concurring with ADAR neighbour preferences [37] (see Figure S1 and 4B). On protease, only the ‘T’ enrichment at the -2 nt (Figure 4B) in agreement to previously reported ADAR neighbor preferences was found.

Sequence analysis of RT p66 ‘A to G’ mutations also showed signatures of ADAR [37], with enrichments of ‘T’ at the -1 over ‘G’ and ‘C’ and -2 positions (enrichment of ‘A’ at the +4 positions ($p < 0.05$).

For RT codon mutated p66 ($p<0.01$), there were enrichments of ‘T’ at the -1, -2 nt, enrichment of ‘A’ bases at the +3 nt, and enrichments of ‘G’ at the +1 and +4 nt positions ($p <0.01$).

For both wild-type and codon mutated p66, the large representation of multiple mutation variants masked single substitution variants with ADAR neighbour preferences found with enrichments of ‘T’ at the -1 and -2 nt positions of the 3’ end (Figure S1).

Effects of 'A to G' Mutations on Protein Thermostability

The distinct observed robustness of HIV to adapt to drugs, immune pressures and the error prone nature of RTs may be attributed to the ability of viral protein structures to buffer fitness by minimizing the change in mutational free energies ($\Delta\Delta G_{Mt}$) [41]. When significantly destabilized (e.g. >4kcal/mol), proteins may adopt a substantially different fold or be misfolded [42]. Taking advantage of relatively large numbers of pooled 'A to G' mutations and other transition and transversion mutations, their distributions in $\Delta\Delta G_{Mt}$ were determined using protein design tools: Rosetta [39] and FoldX [43]. All the mutations were combined to calculate the change in $\Delta\Delta G_{Mt}$ with the threshold of significance for different distributions of $\Delta\Delta G_{Mt}$ at >1 kcal/mol for Rosetta [39] and $\Delta\Delta G_{Mt}$ >0.46 kcal/mol for FoldX [43].

The distribution of $\Delta\Delta G_{Mt}$ of all the mutations in Gag-p66 genes (Protease was excluded as it had no classified hypermutations) in both protein analytical tools showed that most 'A to G' mutations exerted little or negative changes in $\Delta\Delta G_{Mt}$ distributions, while other mutations typically had destabilizing effects (Figure 5A). There was also a secondary peak in the Rosetta results showing a notable distribution of 'A to G' transitions mutants with $\Delta\Delta G_{Mt}$ around a stabilizing -3.99 kcal/mol attributed to Gag (Figure 5A and 5B). From the FoldX result, there were fewer destabilizing mutations with $\Delta\Delta G_{Mt}$ at +5.17 kcal/mol and +7.52 kcal/mol for 'A to G' and other transition and transversion mutations, respectively (Figure 5A). This enforces the trend that 'A to G' mutations tend to elicit less $\Delta\Delta G_{Mt}$, especially since respective secondary peaks were also seen in Gag and the wild-type RT p66 (Figure 5B and C).

The 'A to G' mutations in Gag and RT p66 genes generally had lower $\Delta\Delta G_{Mt}$ compared to other mutations from calculations by both protein design tools (Figure 5A-C). The FoldX results showed significant differences in the distributions of 'A to G' $\Delta\Delta G_{Mt}$ compared to other mutations in both Gag and RT p66. On the other hand, the Rosetta data did not reveal significant $\Delta\Delta G_{Mt}$ differences.

The results suggest that the HIV-1 genes, particularly RT p66 wild-type, were more robust to 'A to G' mutations, especially since 'A to G' mutations in codon mutated p66 were more destabilizing compared to all other mutations; with significant differences in calculations generated by FoldX.

Models of the hypermutations in Gag and wild-type p66 (Table S6) showed the average $\Delta\Delta G_{Mt}$ of multiple mutations as: Gag - 0.37 kcal/mol, 1.75 kcal/mol; p66 - 2.53 kcal/mol, 3.25 kcal/mol as determined using Rosetta and FoldX, respectively. These were in contrast to codon mutated p66 with an average doubling of destabilizing $\Delta\Delta G_{Mt}$ mutants to that of wild-type p66 at 6.48 kcal/mol, 7.00 kcal/mol for Rosetta and FoldX.

Discussion

We set out to study the native mutation rates of HIV-1 RT on HIV-1 genes: Gag, protease and RT p66 subunit in a single replication cycle using our *in vitro* BSL2 assay that is devoid of selection pressures. In the absence of protein translational, immune, viral fitness, and drug selection pressures in our system, the observed biases are presumed to be intrinsic to HIV-RT and that of host cell factors. Although mutations can be contributed from plasmid hosts (DH5α and/or HEK293F), Q5 polymerase and even sequencing artefacts, these were highly unlikely given that the observed phenotypic mutations/bp/replication for *E. coli* and mammalian cells were estimated to be 5.4×10^{-10} and 5.0×10^{-11}, respectively [44], and that Q5 polymerase is one of the most high-fidelity polymerases [45]. Admittedly, there may be a role for primer selection bias when priming the first few bases of the genes, and for this reason, these bases were excluded from analysis and calculation of mutation rates.

While numerous *in vitro* studies have reported the error rates of HIV-1 RT (as previously reviewed in [26,46]), we found only two reports that utilised HIV-1 genes as their template for analysis [47,48]. From our study, we calculated the mutation rates of HIV-1 Gag,

protease, RT p66 and codon mutated p66 to be 4.71 x 10^{-5}, 6.03 x 10^{-5}, 7.09 x 10^{-5} and 1.88 x 10^{-4} mutations/bp (inclusive of hypermutations) respectively, within the range of 1.8 x 10^{-5} – 6.67 x 10^{-4} mutations/bp previously reported [26,46]. Across the HIV-1 genes, there was a predominance of transition mutations consistent across most phyla. This transition bias was hypothesized to be for the conservation of protein functions [49-51]. In our assay, we observed missense mutations to occur at the highest frequency, followed by silent mutations, frameshifts and nonsense. These observations are in agreement with the in-built genetic code probabilities [52] to maintain protein translatability, and avoidance of detrimental effects from frameshifts and nonsense mutations.

The use of different HIV-1 gene templates also elicited varying mutation rates, type of mutations and mutational biases. We noted a general absence of 'A – T' or 'C – G' mutations in the wild-type HIV genes studied, detecting these mutations only in the codon mutated p66 (see Figure 1), suggesting influences that are inbuilt in the sequences and codon usage. This is clearly seen when the codon mutated p66 mutation rate was 2.7 folds higher than the wild-type HIV-1 RT p66. Interestingly, the RNase H domain of the codon mutated HIV-1 RT p66 was also found to have a 7.1 fold increase (3.38 x 10^{-4} mutations/bp) when compared to the wild-type (4.79 x 10^{-5} mutations/bp), possibly due to RNA secondary structural protection/susceptibility [37,53,54].

The increase in 'A to G' hypermutations were observed in HIV-1 Gag, RT p66 and codon mutated RT p66, but not in protease. Possible explanations may be due to the short nucleotide length of protease, thereby avoiding the 'A to G' mutational effects caused by host adenosine deaminases such as double-stranded RNA-specific adenosine deaminase (ADAR). ADAR editing (by either ADAR1 and ADAR2) have been suggested to influence such cell based *in vitro* RT-fidelity assays [55]. ADAR deaminates 'A' residues to inosine (I) residues (read as 'G' by ribosomes due to structural homology), leading to 'C' mis-incorporations in

the plus strand DNA after reverse transcription, resulting in 'G' residues being incorporated during second (minus) stand synthesis [56,57]. ADAR deaminases are believed to be host defence mechanism proteins against viral infections by exerting significant changes on the genomes of such RNA based lifeforms [58-60], exhibiting both pro-viral and anti-viral activities [61-64].

We also found lower occurrences of ADAR neighbour preferences bases in protease (Figure 4) which may explain the lower occurrences of PI resistance in clinical studies [65], alongside better potency than reverse transcriptase inhibitors (RTIs) [66].

The susceptibility of Gag, p66 and codon mutated p66 to 'A to G' hypermutations would result in an accumulation of G bases, that would lead to a translational bias towards glycine, arginine, valine and alanine [52]. This will in turn reduce the occurrences of phenylalanine, isoleucine, tyrosine, histidine and asparagine in these genes. For HIV-1 protease, the mutational bias towards 'A' and 'C' accumulation, which is also the case when 'A to G' hypermutations are not considered for Gag and both p66 genes, would lead to translational bias in the 'A' accumulation towards threonine and lysine, isoleucine, asparagine, arginine and stop codons, while reducing phenylalanine, tryptophan and cysteine occurrences. Similarly, the accumulation of 'C' would lead to proline, serine, leucine, threonine, alanine and arginine while reducing methionine, lysine, glutamic acid, tryptophan, and more importantly, stop codons. It is expected that higher mutations rates can push the virus towards lethal mutagenesis, allowing a possible therapeutic intervention to exploit the ADAR neighbour preference bases flanking target sites.

Structurally, the 'A to G' transitions elicit less destabilization effects in $\Delta\Delta G_{Mt}$ within Gag and wild-type p66. Gag was found to be more robust compared to p66 with its $\Delta\Delta G_{Mt}$ being generally neutral (Rosetta data showed distributions of significantly stabilizing mutations that could have acted as epistatic buffers against destabilizing ones), although this may require

confirmation using cleaved Gag models [67]. On the other hand, codon mutated p66 was less resilient to 'A to G' changes compared to wild-type p66, hinting that RNA viral protein do not just minimize $\Delta\Delta G_{Mt}$ at the structural level as suggested by Tawfik, Berezovsky and colleagues [41], but also at the nucleotide sequence level. With multiple 'A to G' substitutions by ADAR hyper-editing, the average $\Delta\Delta G_{Mt}$ of all p66 mutations (and also Gag) were half that of the codon mutated p66. Lower rates of 'A to G' transition occurred in the wild-type p66 compared to its codon mutated counterpart. Such observations clearly demonstrate that the codon usage of the natural wild-type p66 (and possibly Gag) gene offered some protection against ADAR.

Despite the lack of selection pressure towards functional proteins, we did not find mutations in active sites of the enzymes (protease and p66) nor in the cleavage sites of Gag, with the exception of A431D and F433L in the NC/p1 cleavage site ^{429}RQAN/FLG^{435} of Gag. The detection of non-cleavage mutations in Gag show that such compensation mutations would emerge early in the infection process even in the absence of drugs. While mimiking a single replication cycle, we found known clinical drug resistance mutations (see Table S5 for Gag, protease, and p66). In Gag, the previously reported rare transient mutation E17K (in the matrix), implicated in cytotoxic T lymphocyte (CTL) immune evasion resistance [68,69] was detected. In protease, K70T, a minority mutation associated with resistance to PIs [70] was also found. In p66, a known polymorphic mutation K103R (in the palm domain), which when combined with V179D (not found in our assay), reduced susceptibility to non-nucleoside reverse transcriptase inhibitors (NNRTIs): nevirapine (NVP) and efavirenz (EFV) by about 15 folds [71]. Interestingly. the observed K103R originated from an 'A to G' hypermutation event, specifically from "AAA" to "AGA" suggesting that drug resistance mutations can be induced by ADAR.

Though in much reduced frequency, some mutations were also found in crucial functional sites. In Gag, P222L occurred in the cyclophilin A (CyPA) binding site (along with

G221), potentially affecting the binding of the capsid to CyPA as previously reported for P222A [72]. Given that the probability of P (CCT) to L (CTT) is 0.33 in having a T substitution in the 2nd codon position, and that P (CCT) to A (GCT) is also 0.33 in having the first C mutated to a G, there were equal probabilities. However, when considering that our Gag mutation analysis did not show a transition mutation link between C to G (Figure 1), the clinical mutation is a rare-occurrence (under additional selection pressures like drug) or a result of two or more mutation steps. Mutations A431D in the nucleocapsid and P453T in the p6 domain were also observed, with reported clinical drug resistance counterparts as A431V and a L449F/P453T pair [73]. In p66, the substitution F61S was observed, implicated in strand synthesis with F61Y/L/W altering activity [74]. Similarly, we observed P95L, in which P95 was reported to be at the dimerization interface for the formation of the bottom of NNRTI pocket [75], and a proposed target amino acid in NNRTI design together with N137 and P140 [76]. Although P95 was reported to be in a highly conserved location [75], it was present in our assay with a much limited sample size compared to the virions generated daily in an infected patient.

The occurrence of previously reported *in vitro* and clinical mutations in our assay show that these mutations occurred very early in infection without immune surveillance and drug selection pressures, possibly contributed by HIV gene specific biases (Figure 1). The general avoidance of inducing mutations at crucial sites at the nucleotide stage and protective effects in natural nucleotide codon usage (compared to p66 codon mutated) to minimize changes in protein $\Delta\Delta G_{Mt}$, seems to suggest in-built sequence barriers of self-preservation. Of clinical relevance is that these mutations occurred within our mimic of a single cycle of replication, which when taking into account that HIV-1 generates approximately 10^9 virions per day in an infected individual [77] drug resistance virions would have been made even within the first day of infection.

In the absence of selection pressures, mutations that are selected against can be better detected, allowing for a more comprehensive analysis of the HIV-1 RT bias that can be missed in multiplexing methods. Given that many HIV proteins can function in intense drug/immune selection environments with significant reduced activity [78,79], an all-out campaign against HIV could involve inhibitors against low-frequency occurring drug resistance mutations. This pushes the HIV infection towards Muller's ratchet [80-82] or by increasing viral mutation rate beyond pass a critical threshold towards lethal mutagenesis for error catastrophe [83,84]. Since HIV-1 replicates at the edge of this critical threshold [85], a slight increase in the mutation rate may be sufficient to bring about error catastrophe, which have been successfully induced by mutagenic nucleoside analogues [86-88].

Coupling with *in silico* tools, such as those incorporating probability mutational change, it may be possible to identify mutational hotspots. As it was previously shown that protease and RT drug cross resistance have a structural basis governed by drug resistance mutations [89,90], the bias of restricting specific amino acid changes by the absence of A-T, C-G mutation occurrences can be exploited. However, such an approach will require an in-depth understanding of HIV-RT mutations that are selected against at protein functional levels.

Through the mutational biases and mutation rate of HIV-1 RT, it is possible to calculate the mutational events leading to the zoonotic transmission of SIV to HIV, opening up surveillance of emerging viral threats [2], especially given that RNA viruses are the most likely to species jump [1].

In conclusion, we have established an assay and characterised HIV-1 RT mutations on HIV-1 Gag, protease and RT p66 in a safe BSL2 environment, allowing for insights to the mutational bias and mutation rate of HIV in the absence of biological selection pressures. Through our analysis, we found effects of ADAR in increasing the genetic diversity beyond RT to generate rare and documented clinical drug resistance mutations within a single

replication cycle and thereby very early in infection. The assay can provide deeper insights relevant for drug and vaccine development and be applied for horizontal understanding to other viruses.

Acknowledgments

We thank Chinh Tran-To Su, Wai-Heng Lua and Wei-Li Ling for useful comments and discussion.

Disclosure of interest

The authors report no conflict of interest.

Funding

This work was supported by the A*STAR Industry Alignment Fund (IAF) grant IAF111149 and by EDDC, A*STAR.

Author Contributions

JYY, PY, and GRG performed the *in vitro* experiments. JYY, PY, GRG, and DWSK analysed the results. DWSK performed the *in silico* analysis. JYY, PY, GRG, DWSK, and SKEG discussed and drafted the manuscript. SKEG conceived the idea and supervised the whole project. All authors read and approved the manuscript.

References

1. Woolhouse MEJ, Haydon DT, Antia R. Emerging pathogens: the epidemiology and evolution of species jumps. Trends in Ecology & Evolution. 2005;20(5):238-244.
2. Flanagan ML, Parrish CR, Cobey S, et al. Anticipating the Species Jump: Surveillance for Emerging Viral Threats. Zoonoses and Public Health. 2012;59(3):155-163.
3. Wertheim JO, Worobey M. Dating the Age of the SIV Lineages That Gave Rise to HIV-1 and HIV-2. PLOS Computational Biology. 2009;5(5):e1000377.
4. Zhu T, Korber BT, Nahmias AJ, et al. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature. 1998 1998/02/01;391(6667):594-597.
5. Freed EO. HIV-1 assembly, release and maturation. Nature Reviews. 2015;13:484-496.
6. Pettit SC, Everitt LE, Choudhury S, et al. Initial Cleavage of the Human Immunodeficiency Virus Type 1 GagPol Precursor by Its Activated Protease Occurs by an Intramolecular Mechanism. Journal of Virology. 2004;78(16):8477-8485.
7. Fun A, van Maarseveen NM, Pokorná J, et al. HIV-1 protease inhibitor mutations affect the development of HIV-1 resistance to the maturation inhibitor bevirimat. Retrovirology. 2011;8(1):70.
8. Fun A, Wensing AMJ, Verheyen J, et al. Human Immunodeficiency Virus gag and protease: partners in resistance. Retrovirology. 2012;9(1):63.
9. Clavel F, Mammano F. Role of Gag in HIV Resistance to Protease Inhibitors. Viruses. 2010;2(7):1411-1426.
10. Yu Y, Wang J, Shao Q, et al. Effects of drug-resistant mutations on the dynamic properties of HIV-1 protease and inhibition by Amprenavir and Darunavir. Sci Rep. 2015;5:10517.
11. Wensing AM, Calvez V, Ceccherini-Silberstein F, et al. 2019 update of the drug resistance mutations in HIV-1. Topics in Antiviral Medicine. 2019 Sep;27(3):111-121.
12. Park J, Sayer J, Aniana A, et al. Binding of clinical inhibitors to a model precursor of a rationally selected multidrug resistant HIV-1 protease is significantly weaker than that to the released mature enzyme. Biochemistry. 2016;55:2390-2400.
13. Su CT-T, Kwoh C-K, Verma CS, et al. Modeling the full length HIV-1 Gag polyprotein reveals the role of its p6 subunit in viral maturation and the effect of non-cleavage site mutations in protease drug resistance. Journal of Biomolecular Structure and Dynamics. 2017:1-12.
14. Clutter DS, Jordan MR, Bertagnolio S, et al. HIV-1 drug resistance and resistance testing. Infection, Genetics and Evolution. 2016;46:292-307.
15. Gupta RK, Kohli A, McCormick AL, et al. Full-length HIV-1 Gag determines protease inhibitor susceptibility within in-vitro assays. AIDS. 2010;24(11):1651-1655.
16. Dam E, Quercia R, Glass B, et al. Gag Mutations Strongly Contribute to HIV-1 Resistance to Protease Inhibitors in Highly Drug-Experienced Patients besides Compensating for Fitness Loss. PLoS Pathogens. 2009;5(3):e1000345.
17. Parry CM, Kohli A, Boinett CJ, et al. Gag Determinants of Fitness and Drug Susceptibility in Protease Inhibitor-Resistant Human Immunodeficiency Virus Type 1. Journal of Virology. 2009;83(18):9094-9101.
18. Dam E, Quercia R, Glass B, et al. Gag mutations strongly contribute to HIV-1 resistance to Protease Inhibitors in highly drug-experienced patients besides compensating for fitness loss. Plos Pathogen. 2009;5(3):e1000345.

19. Tamiya S, Mardy S, Kavlick MF, et al. Amino acid insertions near Gag cleavage sites restore the otherwise compromised replication of human immunodeficiency virus type 1 variants resistant to Protease Inhibitors. Journal of Virology. 2004;78(21):12030-12040.
20. Maguire MF, Guinea R, Griffin P, et al. Changes in Human Immunodeficiency Virus Type 1 Gag at Positions L449 and P453 Are Linked to I50V Protease Mutants In Vivo and Cause Reduction of Sensitivity to Amprenavir and Improved Viral Fitness In Vitro. Journal of Virology. 2002;76(15):7398-7406.
21. Gatanaga H, Suzuki Y, Tsang H, et al. Amino Acid Substitutions in Gag Protein at Non-cleavage sites Are Indispensable for the Development of a high Multitude of HIV-1 Resistance against Protease Inhibitors. The Journal of Biological Chemistry. 2002;277(8):5952-5961.
22. Lightfoote MM, Coligan JE, Folks TM, et al. Structural characterization of reverse transcriptase and endonuclease polypeptides of the acquired immunodeficiency syndrome retrovirus. Journal of Virology. 1986;60(2):771-775.
23. Hostomsky Z, Hostomska Z, Fu TB, et al. Reverse transcriptase of human immunodeficiency virus type 1: functionality of subunits of the heterodimer in DNA synthesis. Journal of Virology. 1992;66(5):3179-82.
24. Sarafianos SG, Marchand B, Das K, et al. Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. J Mol Biol. 2009;385(3):693-713.
25. Le Grice SF, Naas T, Wohlgensinger B, et al. Subunit-selective mutagenesis indicates minimal polymerase activity in heterodimer-associated p51 HIV-1 reverse transcriptase. EMBO J. 1991;10(12):3905-3911.
26. Yeo JY, Goh GR, Su CT-T, et al. The Determination of HIV-1 RT Mutation Rate, Its Possible Allosteric Effects, and Its Implications on Drug Resistance. Viruses. 2020;12(3):297.
27. Naldini L, Blömer U, Gallay P, et al. In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. Science. 1996;272(5259):263-267.
28. Chan WT, Verma Chandra S, Lane David P, et al. A comparison and optimization of methods and factors affecting the transformation of Escherichia coli. Bioscience Reports. 2013;33(6):931-937.
29. Ling W-L, Lua W-H, Poh J-J, et al. Effect of VH–VL Families in Pertuzumab and Trastuzumab Recombinant Production, Her2 and FcγIIA Binding. Frontiers in Immunology. 2018;9.
30. Lua W-H, Ling W-L, Yeo JY, et al. The effects of Antibody Engineering CH and CL in Trastuzumab and Pertuzumab Recombinant Models: Impact on antibody production and antigen-binding. Scientific Reports. 2018;8(1):718.
31. Lua W-H, Su CT-T, Yeo JY, et al. Role of the IgE variable heavy chain in FcεRIα and superantigen binding in allergy and immunotherapy. Journal of Allergy and Clinical Immunology. 2019.
32. Sim J-Z, Nguyen P-V, Lee H-K, et al. GelApp: Mobile gel electrophoresis analyser. Nature Methods Application Notes. 2015.
33. Poh JJ, Gan SKE. The Determination of Factors involved in Column-Based Nucleic Acid Extraction and Purification. Journal of Bioprocessing & Biotechniques. 2014;04(03).
34. Koh W-SD, Chan K-F, Wu W, et al. Yet Another Quick Assembly, Analysis and Trimming Tool (YAQAAT): A server for the automated assembly and analysis of Sanger sequencing data. ResearchGate preprint. 2020.

35. Sim JZ, Nguyen PV, Zang Y, et al. DNA2App: Mobile sequence analyser. Scientific Phone Apps and Mobile Devices. 2016;2(1):2.
36. LosAlamos. HIV Sequence Database.
37. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. Nature Communications. 2011;2(1):319.
38. Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics. 2006;22(12):1536-1537.
39. Park H, Bradley P, Greisen P, et al. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. Journal of Chemical Theory and Computation. 2016 2016/12/13;12(12):6201-6212.
40. Williams CJ, Headd JJ, Moriarty NW, et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Science. 2018;27(1):293-315.
41. Tokuriki N, Oldfield CJ, Uversky VN, et al. Do viral proteins possess unique biophysical features? Trends in Biochemical Sciences. 2009;34(2):53-59.
42. Strokach A, Corbi-Verge C, Kim PM. Predicting changes in protein stability caused by mutation using sequence-and structure-based methods in a CAGI5 blind challenge. Human Mutation. 2019;40(9):1414-1423.
43. Delgado J, Radusky LG, Cianferoni D, et al. FoldX 5.0: working with RNA, small molecules and a new graphical interface. Bioinformatics. 2019;35(20):4168-4169.
44. Drake JW, Charlesworth B, Charlesworth D, et al. Rates of spontaneous mutation. Genetics. 1998;148(4):1667-1686.
45. Potapov V, Ong JL. Examining sources of error in PCR by single-molecule sequencing. PloS One. 2017;12(1):e0169774.
46. Menéndez-Arias L. Mutation Rates and Intrinsic Fidelity of Retroviral Reverse Transcriptases. Viruses. 2009;1(3):1137-1165.
47. Geller R, Domingo-Calap P, Cuevas JM, et al. The external domains of the HIV-1 envelope are a mutational cold spot. Nature Communications. 2015 2015/10/09;6(1):8571.
48. Ji J, Loeb LA. Fidelity of HIV-1 Reverse Transcriptase Copying a Hypervariable Region of the HIV-1 env Gene. Virology. 1994 1994/03/01/;199(2):323-330.
49. Lyons DM, Lauring AS. Evidence for the Selective Basis of Transition-to-Transversion Substitution Bias in Two RNA Viruses. Molecular Biology and Evolution. 2017;34(12):3205-3215.
50. Stoltzfus A, Norris RW. On the Causes of Evolutionary Transition:Transversion Bias. Molecular Biology and Evolution. 2016;33(3):595-602.
51. Rosenberg MS. Patterns of Transitional Mutation Biases Within and Among Mammalian Genomes. Molecular Biology and Evolution. 2003;20(6):988-993.
52. Chan K-F, Koukouravas S, Yeo JY, et al. Probability of change in life: Amino acid changes in single nucleotide substitutions. Biosystems. 2020;193-194:104135.
53. Tian N, Yang Y, Sachsenmaier N, et al. A structural determinant required for RNA editing. Nucleic Acids Research. 2011;39(13):5669-5681.
54. Dawson TR, Sansam CL, Emeson RB. Structure and Sequence Determinants Required for the RNA Editing of ADAR2 Substrates. Journal of Biological Chemistry. 2004 February 6, 2004;279(6):4941-4951.
55. Abram ME, Ferris AL, Shao W, et al. Nature, Position, and Frequency of Mutations Made in a Single Cycle of HIV-1 Replication. Journal of Virology. 2010;84(19):9864-9878.
56. Bass BL, Weintraub H. An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell. 1988;55(6):1089-1098.

57. Nishikura K. Functions and Regulation of RNA Editing by ADAR Deaminases. Annual Review of Biochemistry. 2010;79(1):321-349.
58. Di Giorgio S, Martignano F, Torcia MG, et al. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Science Advances. 2020;6(25):eabb5813.
59. Piontkivska H, Frederick M, Miyamoto MM, et al. RNA editing by the host ADAR system affects the molecular evolution of the Zika virus. Ecology and Evolution. 2017;7(12):4475-4485.
60. Piontkivska H, Matos LF, Paul S, et al. Role of Host-Driven Mutagenesis in Determining Genome Evolution of Sigma Virus (DMelSV; Rhabdoviridae) in Drosophila melanogaster. Genome Biology and Evolution. 2016;8(9):2952-2963.
61. Clerzius G, Gélinas J-F, Daher A, et al. ADAR1 Interacts with PKR during Human Immunodeficiency Virus Infection of Lymphocytes and Contributes to Viral Replication. Journal of Virology. 2009;83(19):10119-10128.
62. Doria M, Neri F, Gallo A, et al. Editing of HIV-1 RNA by the double-stranded RNA deaminase ADAR1 stimulates viral infection. Nucleic Acids Research. 2009;37(17):5848-5858.
63. Gélinas J-F, Clerzius G, Shaw E, et al. Enhancement of Replication of RNA Viruses by ADAR1 via RNA Editing and Inhibition of RNA-Activated Protein Kinase. Journal of Virology. 2011;85(17):8460-8466.
64. Weiden MD, Hoshino S, Levy DN, et al. Adenosine Deaminase Acting on RNA-1 (ADAR1) Inhibits HIV-1 Replication in Human Alveolar Macrophages. PLOS ONE. 2014;9(10):e108476.
65. Lathouwers E, Wong EY, Luo D, et al. HIV-1 resistance rarely observed in subjects using darunavir once-daily regimens across clinical studies. HIV Clinical Trials. 2017 2017/11/02;18(5-6):196-204.
66. Titanji BK, Pillay D, Jolly C. Combination antiretroviral therapy and cell–cell spread of wild-type and drug-resistant human immunodeficiency virus-1. Journal of General Virology. 2017;98(4):821-834.
67. Samsudin F, Gan SK-E, Bond PJ. The Structural Basis for Gag Non-Cleavage Site Mutations in Determining HIV-1 Viral Fitness. bioRxiv. 2020:2020.07.05.188326.
68. Sanchez-Merino V, Farrow MA, Brewster F, et al. Identification and Characterization of HIV-1 CD8+ T Cell Escape Variants with Impaired Fitness. The Journal of Infectious Diseases. 2008;197(2):300-308.
69. Sanchez-Merino V, Nie S, Luzuriaga K. HIV-1-Specific CD8+ T Cell Responses and Viral Evolution in Women and Infants. The Journal of Immunology. 2005 2005/11/15/;175(10):6976-6986.
70. Mohamed S, Penaranda G, Gonzalez D, et al. Comparison of ultra-deep versus Sanger sequencing detection of minority mutations on the HIV-1 drug resistance interpretations after virological failure. AIDS. 2014;28(9):1315-1324.
71. Parkin NT, Gupta S, Chappey C, et al. The K101P and K103R/V179D Mutations in Human Immunodeficiency Virus Type 1 Reverse Transcriptase Confer Resistance to Nonnucleoside Reverse Transcriptase Inhibitors. Antimicrobial Agents and Chemotherapy. 2006;50(1):351-354.
72. De Iaco A, Luban J. Cyclophilin A promotes HIV-1 reverse transcription but its effect on transduction correlates best with its effect on nuclear entry of viral cDNA. Retrovirology. 2014 2014/01/30;11(1):11.
73. Banke S, Lillemark MR, Gerstoft J, et al. Positive Selection Pressure Introduces Secondary Mutations at Gag Cleavage Sites in Human Immunodeficiency Virus Type

1 Harboring Major Protease Resistance Mutations. Journal of Virology. 2009;83(17):8916-8924.
74. Fisher TS, Darden T, Prasad VR. Substitutions at Phe61 in the β3-β4 Hairpin of HIV-1 Reverse Transcriptase Reveal a Role for the Fingers Subdomain in Strand Displacement DNA Synthesis. J Mol Biol. 2003 2003/01/17/;325(3):443-459.
75. Ceccherini-Silberstein F, Gago F, Santoro M, et al. High Sequence Conservation of Human Immunodeficiency Virus Type 1 Reverse Transcriptase under Drug Pressure despite the Continuous Appearance of Mutations. Journal of Virology. 2005;79(16):10718-10729.
76. Auwerx J, Van Nieuwenhove J, Rodríguez-Barrios F, et al. The N137 and P140 amino acids in the p51 and the P95 amino acid in the p66 subunit of human immunodeficiency virus type 1 (HIV-1) reverse transcriptase are instrumental to maintain catalytic activity and to design new classes of anti-HIV-1 drugs. FEBS Letters. 2005;579(11):2294-2300.
77. Santoro MM, Perno CF. HIV-1 Genetic Variability and Clinical Implications. ISRN Microbiol. 2013;2013:481314-481314.
78. Chung S, Miller JT, Lapkouski M, et al. Examining the role of the HIV-1 reverse transcriptase p51 subunit in positioning and hydrolysis of RNA/DNA hybrids. Journal of Biological Chemistry. 2013;288(22):16177-16184.
79. Betancor G, Alvarez M, Marcelli B, et al. Effects of HIV-1 reverse transcriptase connection subdomain mutations on polypurine tract removal and initiation of (+)-strand DNA synthesis. Nucleic Acids Research. 2015;43(4):2259-2270.
80. Hu W-S, Temin HM. Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. Proceedings of the National Academy of Sciences. 1990;87(4):1556-1560.
81. Yuste E, Sánchez-Palomino S, Casado C, et al. Drastic fitness loss in human immunodeficiency virus type 1 upon serial bottleneck events. Journal of Virology. 1999;73(4):2745-2751.
82. Ormond L, Liu P, Matuszewski S, et al. The combined effect of oseltamivir and favipiravir on influenza A virus evolution. Genome biology and evolution. 2017;9(7):1913-1924.
83. Summers J, Litwin S. Examining The Theory of Error Catastrophe. Journal of Virology. 2006;80(1):20-26.
84. Bull JJ, Sanjuán R, Wilke CO. Theory of Lethal Mutagenesis for Viruses. Journal of Virology. 2007;81(6):2930-2939.
85. Drake JW, Holland JJ. Mutation rates among RNA viruses. Proceedings of the National Academy of Sciences. 1999;96(24):13910-13913.
86. Dapp MJ, Clouser CL, Patterson S, et al. 5-Azacytidine Can Induce Lethal Mutagenesis in Human Immunodeficiency Virus Type 1. Journal of Virology. 2009;83(22):11950-11958.
87. Harris KS, Brabant W, Styrchak S, et al. KP-1212/1461, a nucleoside designed for the treatment of HIV by viral mutagenesis. Antiviral Research. 2005;67(1):1-9.
88. Loeb LA, Essigmann JM, Kazazi F, et al. Lethal mutagenesis of HIV with mutagenic nucleoside analogs. Proceedings of the National Academy of Sciences. 1999;96(4):1492-1497.
89. Chiang Ron Z-H, Gan Samuel K-E, Su Chinh T-T. A computational study for rational HIV-1 non-nucleoside reverse transcriptase inhibitor selection and the discovery of novel allosteric pockets for inhibitor design. Bioscience Reports. 2018;38(2):BSR20171113.

90. Su CT-T, Ling W-L, Lua W-H, et al. Structural analyses of 2015-updated drug-resistant mutations in HIV-1 protease: an implication of protease inhibitor cross-resistance. BMC Bioinformatics. 2016 2016/12/22;17(19):500.

Table 1: Calculated error rates of HIV-1 RT on the respective HIV-1 genes

HIV-1 Gene	No. of clones	Nucleotide Length	Total no. of bases	Excluding 'A to-G' Hypermutations		Including 'A to G' Hypermutations	
				No. of mutations	Mutation rate (mutations/bp)	No. of mutations	Mutation rate (mutations/bp)
Gag	801	1485	1,189,485	40	3.36×10^{-5}	56	4.71×10^{-5}
protease	640	285	182,400	11	6.03×10^{-5}	-	-
RT p66	571	1680	959,280	41	4.27×10^{-5}	68	7.09×10^{-5}
Codon mutated RT p66	700	1617	1,131,900	74	6.53×10^{-5}	213	1.88×10^{-4}

Mutation rates were calculated as the ratio of the total number of mutations and the total number of nucleotide bases. The rates were within the reported range of 1.8×10^{-5} – 6.67×10^{-4} as previously reviewed [26,46].

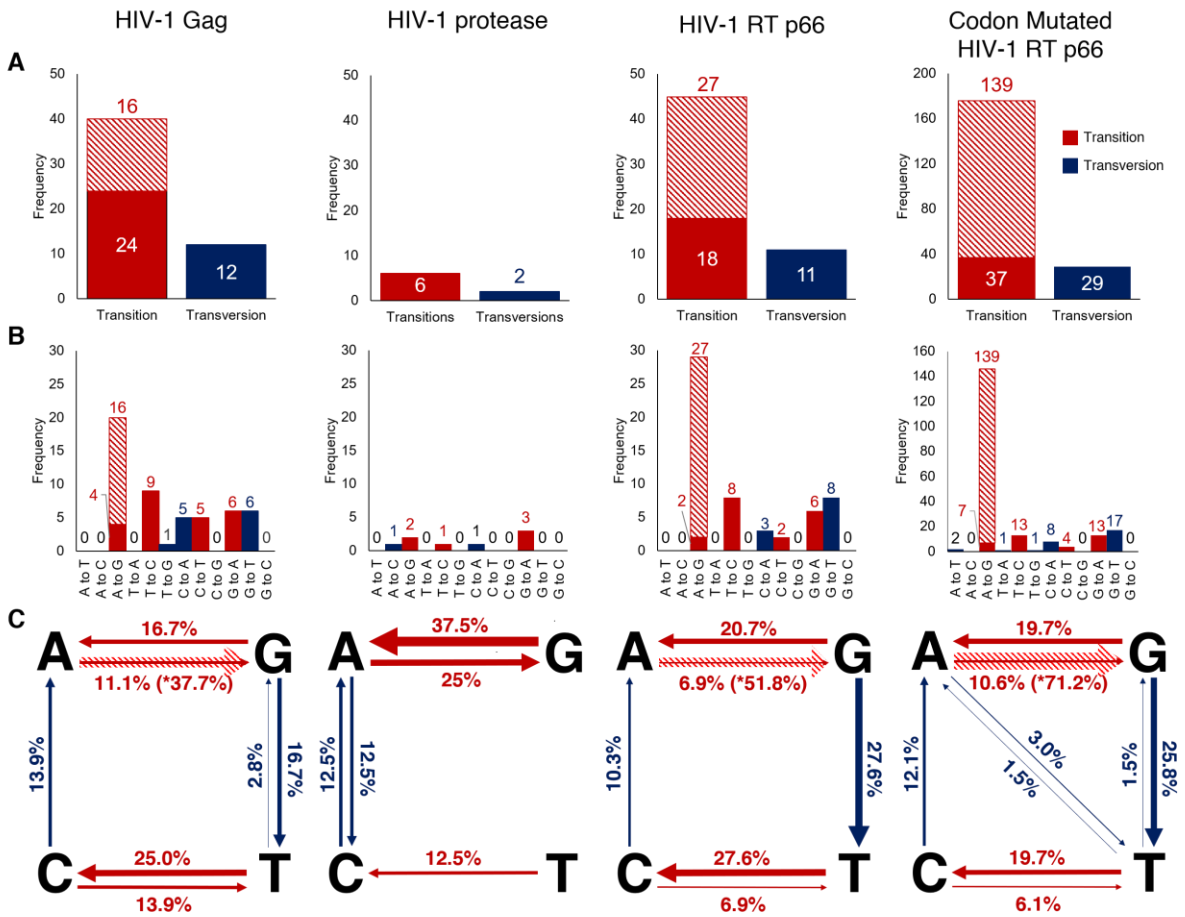


Figure 1. Generated HIV-1 Gag, protease, RT p66 and Codon Mutated RT p66 nucleotide substitution mutations using HIV-1 RT.

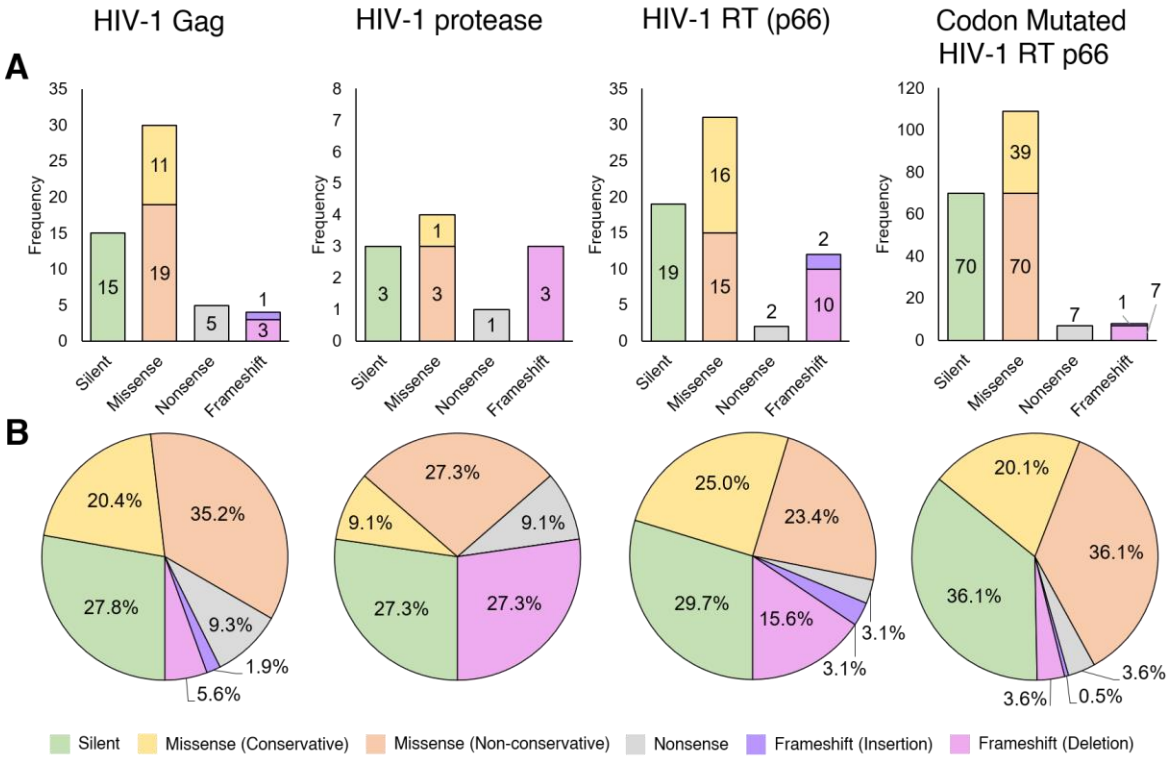


Figure 2. Generated amino acid mutations in HIV genes using HIV-1 RT.

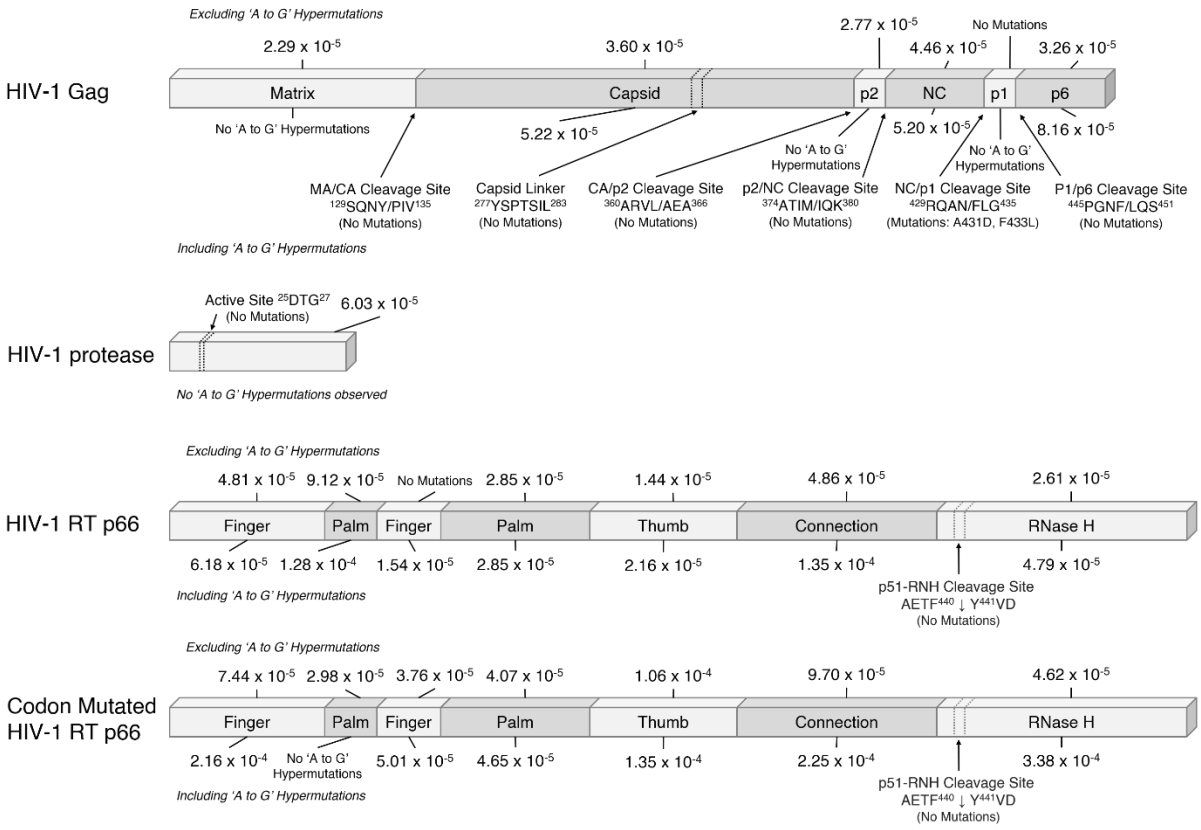


Figure 3. Mutation rates (mutations/bp) calculated in the domains of HIV-1 Gag, protease, RT p66 and Codon Mutated RT p66.

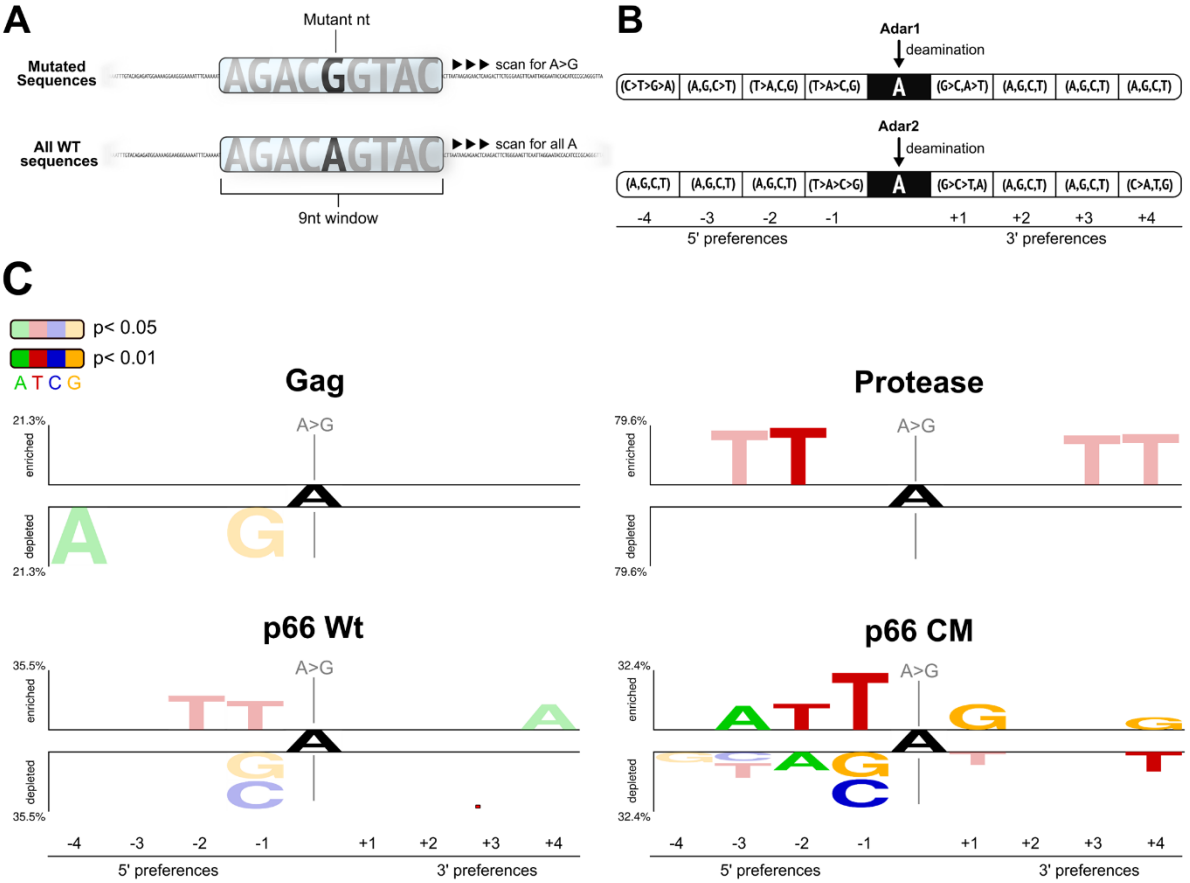
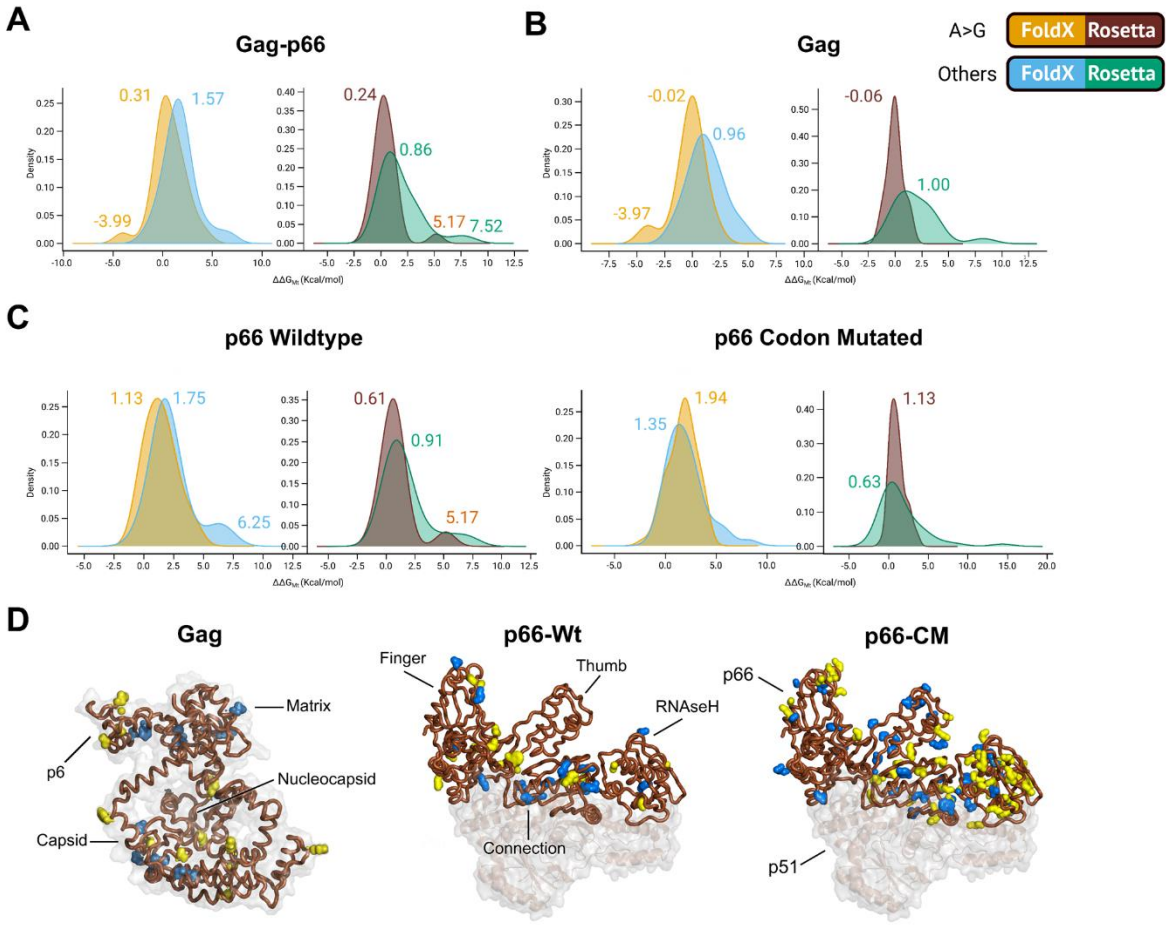


Figure 4. Two Sample logo analysis of HIV-1 Gag, protease, RT p66 (p66 Wt) and p66 Codon Mutated RT (p66 CM)



705

706 **Figure 5. Distributions of the change in mutational free energies ($\Delta\Delta G_{Mt}$) of pooled**

707 **mutations.**

Figure Captions

Figure 1. Generated HIV-1 Gag, protease, RT p66 and Codon Mutated RT p66 nucleotide substitution mutations using HIV-1 RT.

(A) Bar chart of the transitions and transversions in HIV-1 Gag (n = 52), protease (n = 8), RT p66 (n = 56) and codon mutated p66 (n = 205). Transitions and transversions are shown in red and blue, respectively. 'A to G' hypermutations are shown separately in diagonal stripes. (B) Bar chart showing substitution mutations in HIV-1 Gag, protease, RT p66 and codon mutated p66 sequences. (C**)** Schematic diagram of nucleotide substitutions on the respective HIV-1 genes (expressed as a percentage).

Figure 2. Generated amino acid mutations using HIV-1 RT.

(A) Bar chart of amino acid mutations observed in HIV-1 Gag (n = 54), protease (n = 11), RT p66 (n = 64) and codon mutated RT p66 (n = 194). (B) Pie chart of the percentage proportions of the type of translation effects resulting from the mutations. For the codon mutated RT p66, two variants (20 and 80 a,b) containing deletions without frameshifts are not shown. Silent, missense (conservative), missense (non-conservative), nonsense, frameshift (insertion) and frameshift (deletion) mutations are shown in green, yellow, orange, grey, purple and pink respectively.

Figure 3. Mutation rates (mutations/bp) calculated in the domains of HIV-1 Gag, protease, RT p66 and Codon Mutated RT p66.

Mutation rates calculated as the total number of mutations and the total number of nucleotide bases were counted separately excluding and including A to G hypermutations. Truncations that span across multiple domains were excluded in the calculations.

Figure 4. Two Sample logo analysis of HIV-1 Gag, protease, RT p66 (p66 Wt) and Codon Mutated RT (p66 CM)

(A) Underlying neighbor preferences 4 nt upstream and downstream of all 'A to G' mutations were determined against all possible background mutations from a 9 nt sliding window. (B) A re-representation of previous reports [37] on the neighbour preferences of ADAR1 and 2 isoforms. Base preferences are shown with inequality signs, where the identity of the base did not significantly contribute to the editing of the 4 bases are delimited by commas. (C) Two Sample Logos showed enrichment and depletion of neighbouring nucleotides flanking A to G mutations. Bases were coloured opaque when $p < 0.01$, and translucent when $p < 0.05$. Statistics: two sample t-test without Bonferroni correction.

Figure 5. Distributions of the change in mutational free energies ($\Delta\Delta G_{Mt}$) of pooled mutations.

Pooled $\Delta\Delta G_{Mt}$ of point mutations from (A) Gag-p66, (B) Gag (C) wild-type RT p66 (p66-Wt) and codon mutated RT p66 (p66-CM) genes. (D) Positions of 'A to G' mutations in the protein structures of Gag and p66 are shown in yellow while all other mutations are in blue. There were both 'A to G' and 'other' mutations at position 462 of p66-CM labelled blue. The C and N terminals and locations of major domains and functional protein regions are shown with labels. Peaks are numbered with corresponding $\Delta\Delta G_{Mt}$ values. The mutational free energies were modelled with the Rosetta Cartesian_ddg and FoldX BuildPDB protocols. $\Delta\Delta G_{Mt}$ values or $\Delta\Delta G_{Mt}$ differences between different distributions of >1 kcal for Rosetta (29) and ΔΔGMt >0.46 kcal/mol for FoldX (30) were considered to be significant. Note that Gag exists in both the Compact and Extended states and since the former precedes the later during viral assembly, the former was used.