Vulnerable targets in HIV-1 Pol for attenuation-based vaccine design

Doty B.A. Ojwach[a], Paradise Madlala[a], Michelle Gordon [b], Thumbi Ndung'u [a, c, d, e, f], Jaclyn K. Mann[a*]

[a] HIV Pathogenesis Programme, Doris Duke Medical Research Institute, University of KwaZulu-Natal, Durban, South Africa; [b] KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Department of Laboratory Medicine & Medical Sciences, University of KwaZulu-Natal, Durban, South Africa [c] Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA; [d] Africa Health Research Institute, Durban, South Africa; [e] Max Planck Institute for Infection Biology, Berlin, Germany; [f] Division of Infection and Immunity, University College London, London, UK.

*Corresponding author

Jaclyn K. Mann, Ph.D.

HIV Pathogenesis Programme, Doris Duke Medical Research Institute

Nelson R Mandela School of Medicine

University of KwaZulu-Natal

719 Umbilo Road

Congella, Durban 4001

South Africa

Email: mannj@ukzn.ac.za

Highlights

- Reverse transcriptase (RT) and integrase mutations associated with cytotoxic T cell (CTL) escape were confirmed by mutagenesis experiments to decrease viral replication capacity

- RT mutations associated with CTL escape linked with reduced reverse transcription ability
- These fitness costly mutations represent potential vulnerable targets in Pol for an attenuation-based HIV vaccine
- Structural models provide insight into how these RT mutants affect amino acid interactions in the helix clamp, primer grip and catalytic site regions of RT.

Keywords
HIV-1 RT-integrase, replication capacity, site-directed mutagenesis, sequence modelling

Abstract

Identification of viral immune escape mutations that compromise HIV's ability to replicate may aid rational attenuation-based vaccine design. Previously we reported amino acids associated with altered viral replication capacity (RC) from a sequence-function analysis of 487 patient-derived RT-integrase sequences. In this study, site-directed mutagenesis experiments were performed to validate the effect of these mutations on RC. Viral reverse transcripts were measured by quantitative PCR and structural modelling was performed to gain further insight into the effect of reverse transcriptase (RT) mutations on reverse transcription. RT-integrase variants in or flanking cytotoxic T cell epitopes in the RT palm (158S), RT thumb (241I and 257V) and integrase catalytic core domain (124N) were confirmed to significantly reduce RC. RT mutants showed a delayed initiation of viral DNA synthesis. Structural models provide insight into how these attenuating RT mutations may affect amino acid interactions in the helix clamp, primer grip and catalytic site regions.

Introduction

CD8+ cytotoxic T cells (CTL) are known to have a central role in controlling HIV-1 replication, through recognition of viral peptides bound to human leukocyte antigen class I (HLA-I) molecules on the surface of infected cells [1-5]. While HIV-1 readily acquires mutations that affect viral peptide and HLA-I binding as a counter strategy to escape surveillance of CTLs [6], for a subset of these escape mutations the advantages of CTL escape for the virus are offset by costs to viral replication capacity (RC) [7-9]. Indeed, HIV-1 elite controllers tend to target multiple constrained regions of Gag where immune escape has large costs to viral RC, thereby allowing these individuals to avoid immune escape by the virus and to maintain immune pressure [10]. Viral attenuation as a result of immune-driven escape mutations has been proposed as a concept in rational vaccine design where the goal is to focus CTL responses on the most constrained regions of the virus [8, 11]. The most vulnerable regions of the virus cannot be predicted by sequence conservation alone [11-13]. Predictive tools based on higher order sequence analysis [10, 14, 15] and protein structure [11], as well as functional assays [9, 16-18], may inform this vaccine strategy, which should be broad enough to be applicable to the entire population regardless of HLA-I profile [14].

Gag is the best characterised HIV-1 protein in terms of fitness constraints, while the extent to which host immune pressure on other proteins can affect RC of naturally occurring HIV-1 sequences remains limited. The highly conserved Pol protein [19] may contain promising targets for vaccine design [20-23]. While there are fitness-costly drug resistance mutations described in Pol [24-27], there is limited information on the impact of CTL escape mutations in Pol that have a fitness cost [28-30], particularly for HIV-1 subtype C, the most abundant clade in the pandemic. We previously constructed over 400 recombinant viruses encoding patient-derived subtype C reverse transcriptase (RT)-integrase sequences and measured their replication capacities in a green fluorescent protein (GFP) reporter T cell line [31]. Through

this functional analysis, we identified likely CTL escape mutations in RT and integrase that were statistically associated with decreased RC [31]. In the present study, site-directed mutagenesis was performed to confirm their effect on RC. RT activity assays, as well as structural modelling, was performed to gain further insight into their effect on Pol function.

Materials and methods

Site-directed mutagenesis

Site-directed mutagenesis experiments were performed, as previously described [17], to directly confirm the replicative fitness consequences of the RT (E6K, A158S, V241I, I257V, P272K and E297K) and IN (V112I, A124N, Q136K, I201V and I208X) mutations that were statistically associated with decreased viral RC and were in or adjacent to CTL epitopes [31] (Table 1). Mutations at RT codon 272 are frequent and, although only the rare variant 272 K was statistically associated with reduced RC, the most common mutation 272A was also therefore included. Since the consensus 208I in IN was associated with increased RC, two equally frequent mutations – 208L and 208M - at this codon were tested. Although variants at IN codon 119 were only associated with increased RC in our previous analysis (Table 1), 119 variants were previously reported to decrease RC [32] and influence disease progression [33], therefore site-directed mutagenesis was also performed to test the functional effect of variants at this codon.

Table 1: Mutations in HIV-1 RT and IN identified by statistical analysis to alter viral replication capacity [a]

Protein	Codon	AA	RC +AA	RC -AA	*p*-value	*q*-value	CTL epitope [c]	CTL escape[d]
RT	E6	K	0.87	0.92	0.017	0.34	IL8 (B*40:01); SL10 (B*8101)	6K, SL10 (B*81:01) [34]
RT	A158	S	0.87	0.92	0.004	0.21	LA10 (B*54:01); SM9 (B*07, B*8101, B*39); AK9 (A*03:01/A*11:01)	
RT	V241	I	0.82	0.92	0.005	0.19	Flank[e] IW9 (B*5701)	
RT	I257	V	0.80	0.92	0.006	0.21	Flank[e] IW9 (B*5701) and LY12 (B*1501)	
RT	P272	K	0.84	0.92	0.010	0.23	QR9 (A*03:01), YL9 (B*42:01, B*57, B*81)	272A,S,Q, YL9 (B*42:01) [35]
RT	E297	K	0.80	0.92	0.003	0.21	IL9 (B*35:01), IL9 (B*51:01)	297A,G,Q,V, IL9 (B*07) [36]
IN	V112	I	0.88	0.92	0.037	0.36	Flank[e] HF8 (C*05)	
IN	S119	T	0.96	0.92	0.018	0.32	HF8 (C*05), Flank[e] SW10 (B*57)	
IN	S119	P	0.94	0.92	0.032	0.35		
IN	A124[b]	N	0.78	0.92	0.002	0.2	SW10 (B*57)	124N, SW10 (B*57) [37]
IN	Q136	K	0.88	0.92	0.028	0.33	IY9 (B*15:03, B*39)	
IN	I201	V	0.83	0.92	0.040	0.36	GI8 (B*40:02)	
IN	I208	I	0.92	0.86	0.009	0.22	IK9 (A*11)	

[a] These mutations were identified following a sequence-function analysis of patient-derived Pol sequences from a large cohort of chronically infected individuals [31].

[b] This mutation was significantly associated with reduced RC in sequences derived from early infection only [31].

[c] CTL epitope list as available from http://www.hiv.lanl.gov [38].

[d] Confirmed and documented CTL escape as reported in published literature.

[e] The mutation is within 5 amino acids of the optimal CTL epitope.

RT-reverse transcriptase, IN-integrase, RC-replication capacity, CTL-cytotoxic T-cell lymphocyte, AA-amino acid. All associations for n > 5, *p*- value < 0.05 and *q*-value < 0.4.

Mutations of interest were introduced into a patient-derived subtype C RT-integrase sequence (SK-278; GenBank accession number MH487306.1) that was the most similar (98.2% amino acid sequence similarity) of the patient-derived sequences in our previous study [31] to the consensus C RT-integrase sequence [38]. RT-integrase was PCR amplified from the HIV-1 RNA derived from patient SK-278 [31] and cloned into a TOPO plasmid vector using the

TOPO TA 3.1 cloning kit (Invitrogen, Life Technologies Corporation, Carlsbad, USA) according to manufacturer's instructions. Mutations were introduced into the SK-278 RT-integrase sequence inserted in the TOPO plasmid using the QuikChange II XL Site-Directed Mutagenesis kit (Agilent Technologies, Texas, USA) and custom-designed mutagenic primers. The mutant plasmids were then sequenced as previously described [31] to confirm the presence of the desired mutations.

Generation and RC measurement of mutant viruses

Generation and RC measurement of viruses was performed as previously described [31, 39] Briefly, RT-integrase was PCR amplified from the mutant plasmids using the Takara Ex Taq® Hot Start kit (Takara, Shiga, Japan) with 100mer primers that match NL4-3 on either side of *RT-integrase,* to enable the homologous recombination of the PCR product and the pNL4-3Δ*RT-integrase* backbone sequence (in the backbone, *RT-integrase* was deleted and replaced with a BstEII site [39, 40]). CEM-GXR cells were transfected with BstEII-linearized pNL4-3Δ*RT-integrase* plasmid mixed with the mutant *RT-integrase* PCR product via electroporation: homologous recombination of insert and vector in the cells produces infectious recombinant HIV-1. CEM-GXRs are a CEM-derived T cell line which contain a HIV-1 Tat-dependent GFP expression system [41] and express the CD4, CXCR4 and CCR5 receptors [42]. GFP emits green light on exposure to blue light at 488 nm which enables monitoring of infected cells (GFP-positive cells) using a flow cytometer. Cultures were monitored by flow cytometry and when approximately 25% of cells in the culture expressed GFP, virus-containing culture supernatants were harvested. The multiplicity of infection (MOI) of the virus stock was calculated by infecting 1 x 10^6 GXR cells with 400 µl virus stock and quantifying GFP expression at 48 hours post-infection. The titre results were used to calculate the volume of virus stock to infect GXR cells at a MOI of 0.003 for the RC assays as follows: volume of

virus stock = (0.3/proportion GFP-expressing cells) x 400 µl. Subsequently, 1 x 10^6 GXR cells were infected at a MOI of 0.003 and the percentage of infected cells was monitored by flow cytometry on days 2-6 post-infection. The RC was determined by calculating the slope of exponential growth from days 3-6 using the semi-log method in Excel, and normalising this value to the slope of growth of the wildtype SK-278, which was assigned an RC value of 1. RC experiments were performed independently in triplicate and the data averaged.

Quantification of RT activity by real-time PCR

To assess the functional significance of the RT mutations associated with escape on the reverse transcription process, viral DNA transcripts were quantified following infection of GXR cells with wildtype and mutant viruses. Briefly, GXR cells (1.0 x 10^6 cells per well in a 24-well plate) were infected with 1000 ng p24 equivalent of virus in 800 µl of R10 medium and incubated for 2 hours at 37 °C and 5% carbon dioxide. The p24 concentration of virus stocks was determined using HIV-1 p24 Antigen ELISA (Zeptomatrix corporation Buffalo, New York) according to manufacturer's instructions. At 2 hours post infection, the cells were washed twice and re-suspended in 800 µl fresh R10 medium. At 6 and 12 hours post infection, the cells were collected by centrifugation. The cell pellets obtained at different time points were lysed as described previously [43].

To quantify late reverse transcription products, primers (MH531: 5'-TGTGTGCCCGTCTGTTGTGT-3' and MH532: 5'-GAGTCCTGCGTCGAGAGAGC-3') and a probe (LRTP: 5'-(FAM)-CAGTGGCGCCCGAACAGGGA-(TAMRA)-3') were utilized to amplify a region of the cDNA between the 5' LTR sequence and the 5' end of the *gag* gene from the cell lysates [44, 45]. To standardize according to the amount of cellular DNA in the cell lysate, the 18S rRNA housekeeping gene was amplified in each sample using the 18S rRNA endogenous control FAM™ / MGB Probe, non-primer limited (Applied Biosystems, Foster City, California), and the reverse transcript quantity was normalized to the

level of 18S in the lysate. The eukaryotic 18S rRNA allows relative gene expression quantification in cDNA samples when used with other gene expression assays [46].

For each real-time PCR analysis, a standard curve for relative quantification of the late reverse transcripts and 18S was prepared independently by serial dilution (1.6 to 1000 copies) of the lysate from the NL4-3 control infection. The late RT reactions contained 1× TaqMan universal master mix (Applied Biosystems), 10 µM forward primer, 10 µM reverse primer, 2 µM probe primer and 5 µl of template DNA in a 20 µl volume. The 18S reaction contained 1× TaqMan universal master mix, 2 µM 18S RNA and 5 µl of template DNA which was made up to 20 µl total volume with ddH_20. After initial incubations at 50 °C for 2 min and 95 °C for 10 min, 40 cycles of amplification were carried out at 15 s at 95 °C followed by 1 min at 60 °C. Real time analysis was performed using the LightCycler® 480 System (Roche Diagnostics, Risch-Rotkreus, Switzerland). Three independent infections were carried out and real time PCR was performed in duplicate for all the infections.

Modelling of RT mutations

To explore structural changes that could be promoted by the mutations under study, homology modelling of wildtype SK-278 RT-integrase and mutant versions was performed using SWISS-MODEL [47]. The visualization and analysis of the molecular structures was done using UCSF Chimera software [48]. Interactions with the mutant or wildtype residues within 5Å were considered and the changes in the distances of these interactions as a result of the mutation were calculated. Changes in the distances of the RT residues of interest to the nucleic acid template when comparing mutant and wildtype models were explored.

Statistical analysis

ANOVA with Tukey post hoc tests were used to test for differences in RC or RT activity between viruses encoding the mutations and the respective wildtype virus. Statistical tests were performed using Prism 5.0 (GraphPad Software, Inc.).

Results and Discussion

The effect of RT mutants on replication capacity

Using our combined genotype and RC dataset we previously identified specific mutations in RT-integrase, several of which are likely immune-driven, that were associated with decreased viral RC [31] (Table 1). Mutagenesis experiments were performed to confirm and directly test the effects of these RT-integrase mutations on viral RC. The titre data and raw RC data used to calculate RC of the mutant viruses are shown in the supplementary information (Table S1 and Figure S1). The A158S, V241I, I257V and P272K mutations in RT were confirmed to significantly reduce RC (ANOVA with Tukey post hoc tests; $p<0.01$), while E6K, P272A and E297K RT mutations were not significantly different from the wildtype (Figure 1).

The A158S mutation is a variant of the SM9 CTL epitope restricted by the B7 supertype alleles B*07, B*39 and B*81, and it is located in the palm on the β8–α helix E loop which interacts with the substrate [49]. Sequences harbouring V241I (RT) and I257V (RT) exhibited the lowest RC; both of these flank the B*57-restricted epitope IW9 and are in the thumb domain which forms the base of the nucleic acid binding cleft that contributes to RNA/DNA binding [50]. Similarly, the rare P272K variant in the thumb region was confirmed by mutagenesis to lower RC, however RT codon 272 is not well-conserved and common escape mutations at this codon, including P272A, did not alter replication capacity, suggesting that this residue is not an ideal target for attenuation-based vaccine design. In summary, mutations 158S, 241I and

257V at conserved residues in RT were confirmed by mutagenesis to significantly reduce viral RC and represent vulnerable residues in RT.

The effect of RT mutants on RT activity

To gain further mechanistic insights into these fitness-reducing mutations, we measured their reverse transcription ability. A real time PCR assay was utilised that detects only DNA forms that have completed the two template switches of reverse transcription. Viral transcripts measurements from independent infection experiments correlated well (Pearson's correlation, $r > 0.96$ and $p < 0.0001$). At 6 hours post infection, the process of reverse transcription was noticeably delayed or hampered in the 241I and 257V mutants compared to the wildtype virus, and to a lesser extent for the 158S mutant, although this only reached statistical significance for 257V (ANOVA with Tukey post hoc tests; $p < 0.05$) (Figure 2A). However, at 12 hours post infection this signal was not observed (ANOVA, $p < 0,75$; Figure 2B) likely as the peak in late reverse transcripts was reached earlier for the wildtype and there is an overall decline in reverse transcripts by this time point.

Modelling of RT mutants

To further explore the mechanisms underlying the functional costs incurred by 158S, 241I and 257V, structural modelling was performed using the wildtype and mutant RT sequences and changes in the structure were investigated. Mutant 158S is in the RT palm while 241I and 257V are in the RT thumb, as depicted in Figure 3A.

I257V

The thumb is vital for holding the primer/template in position during polymerization and has two antiparallel α-helices, α-H (Asn255 to Ser268) and α-I (Gln278 to Thr286) [50], with the primary sequence (Val254 to Ala288) forming the so called "helix clamp"[51] (Figure 3A and B). As shown in the models in Figure 3B, codon 257 in α-H interacts with both codon 279 and 283 in α-I at less than 5Å, and with the nucleic acid template at 8.66 Å, and the distance of

these interactions with 257 is altered by the isoleucine to valine mutation at this codon (Supplementary Table 2). With the mutation to valine in this structural model, residue 257 is 0.13Å further from the template. The models show a shorter distance between mutant 257V and residue 279 compared to the wildtype 257I (by 0.04 Å), which could explain the longer distance interaction (by 0.1 Å) between mutant 257V and residue 283, as well as the increased distance of mutant 257V from the template. These alterations may partly underlie the decreased RT function attributed to I257V. Of note, the mutant model did not show altered distance between the template and key residues 258, 262 and 266, which were the α-H residues previously reported to play an important role in template-primer binding [52, 53].

A previous steady state analysis of alanine mutants in α-H, revealed that the mutant I257A significantly reduced the efficiency of nucleotide incorporation [52]. Another study of the mutation I257T showed that this mutation made RT more susceptible to protease cleavage, which altered the stability of RT resulting in reduced RC [54]. Interestingly, valine and threonine are nearly of similar shape and volume, with the exception of the methyl substituent in valine replacing the hydroxyl group in threonine, suggesting that 257V may similarly affect RT stability.

V241I

Mutant V241I is flanking the primer grip region (residues 227-235) of the thumb that forms the β12-β13 hairpin of the RT molecule responsible for upholding the primer terminus in a suitable orientation for nucleophilic attack on the incoming deoxy nucleoside triphosphate [55] (Figure 3A and C). In our modelling analyses, the mutant 241I forms several closer interactions (by 0.14, 1.44 and 1.06Å) with residues 266 and 267 (in the α-H helix), as well as residue 269, in comparison to the wildtype 241V (Figure 3C). The model indicates that the mutant decreases the distance of residue 241 interaction with the α-H helix (Supplementary Table 2). While residue 266 is key in binding to the template [52, 53], the altered interaction between 241 and

266 in the mutant versus wildtype models did not affect the distance of the interaction of 266 with the template, however the mutant 241I residue was further (by 0.21 Å) from the template than the wildtype 241V in these models. The mutant model also showed altered orientation of 241I with respect to the nearby (within 5Å) residue 232 in the primer grip (with no influence on the distance of interaction between 232 and the template), which suggests a change in interaction of the mutant 241I with the primer grip. Residue 232 is within 5Å of 241I and different molecules in the side chains of these amino acids have increased (0.46 Å) or decreased distance (0.21 Å) from each other when comparing the mutant and wildtype models. Interestingly, in a previous study of viruses with RT mutations within and flanking the primer grip domain (M230A, L234D, and W239A), the reverse transcription process, proviral DNA synthesis and virion maturation were impaired [56], indicating that certain mutations in this region could also affect the processing of the Gag-Pol polyprotein to achieve virion maturation.

A158S

A158S is in the β8–α helix E (residues 155–174) loop at the junction of the fingers and palm subdomains of HIV-1 RT, and is proximal to the catalytic site (residues 91-119 and 151-157) (Figure 3A and D). In the mutant model, 158S directly forms a hydrogen bond with G155 (Figure 3D) which is part of the catalytic site and this could partly explain the decrease in RT activity of this mutant [57]. Additionally, the mutant 158S model shows altered interaction between 158S and the β5a strand (residues 84-92) (Figure 3D), which has been reported to assist in clamp and strand transfer activities of RT [58]. Specifically, the mutant gains a closer interaction with residues 87 (by 1.48Å) and 89 (by 0.08Å), while its distance from residue 90 is slightly increased (by 0.03Å; Supplementary Table 2). Residue 158 has previously been characterized as having functional importance, where non-conservative substitutions but not all conservative changes negatively affect RT activity while all substitutions at this codon are wildtype for protein stability [57]. While mutant A158T does not affect clamp and strand

transfer activity of HIV-1 RT [58], our modelling analyses indicate that 158T differs from 158S in its orientation and interactions within the clamp activity region of the RT (data not shown). A158S, which is the only mutation observed at this codon in the subtype C sequences studied here [31] appears to be a conservative substitution where the negative effect on viral RC is of a lower magnitude than for mutations 257V and 241I.

The effect of IN mutants on replication capacity

The IN non-consensus variants at codons 112, 124, 136, 201 and 208 were previously statistically associated with decreased RC [31] and were tested for their effect on RC by site-directed mutagenesis in the present study. Only A124N ($p< 0.001$) and I208M ($p<0.01$) had statistically significant lower RC in comparison to the wildtype.

The A124N mutation was associated with the most pronounced decrease in RC of the IN mutations tested here. A124N is in the B*57 restricted SW10 CTL epitope and in a genome-to-genome analysis of HIV-1 subtype B chronic infection, it was identified as one of the most biologically relevant interactions between the host and the virus where individually it was associated with a modest negative change in viral load [59]. It is in the catalytic core domain (CCD) which comprises residues 56–186 that are vital for specific viral DNA binding [60, 61]. The structural effect of the A124N mutation was previously analysed and this mutation was reported to perturb the interaction between the CCD and carboxyl-terminal domain (CTD)[62]. Further, A124N was reported to be resistant to proposed allosteric inhibitors of integrase (ALLINI) [62, 63], which is consistent with the replication defect observed in this study. However, the A124N mutation was statistically associated with decreased RC in our recent infection cohort but not in the chronic infection cohort (Table 1), suggesting that compensation may mask the fitness defect by the chronic infection stage.

The methionine mutation at the conserved codon 208 in the A*11-restricted IK9 epitope was associated with significantly decreased replication capacity, which is consistent with the observation that mutations seldom occur in this epitope-rich region in the CTD (which is necessary for binding of the host DNA [60]) and the reported resistance of I208M to integrase inhibitors [63]. However, another mutation observed at equal frequency at this codon 208L was not significantly different in RC from the wildtype, indicating that the viral fitness effect of mutating this codon is not consistent.

Variants at integrase codon S119 which is also in the CCD domain were significantly associated with altered RC in both subtype B and in our analysis of subtype C [31]; however, whereas the subtype B study identified 119R as being detrimental to RC [32], 119T and 119P were associated with increased RC in subtype C (Table 1) [31]. We therefore tested the effect of S119T, R and P in both HIV-1 subtype C (wildtype SK-278 RT-integrase) and subtype B (NL4-3) backbones. In the SK-278 backbone, only the S119T and S119R grew in culture and were not associated with altered RC (Figure 4). In the NL4-3 backbone, S119T and S119P mutant viruses were successfully harvested, however they had no significant impact on RC when compared to the wildtype (RC of 1.06 and 1.00, respectively, with respect to wildtype NL4-3), while the plasmid encoding S119R did not yield a viable virus in culture, consistent with the reported attenuating effect of this mutation in HIV-1 subtype B infection [32]. Interestingly, S119G, which was rare (0.05% frequency) in our patient cohort, not identified to alter RC in our statistical analyses [31] and not tested by mutagenesis in the present study, was previously linked to rapid disease progression in HIV-1 subtype C chronic infection [33]. Taken together, the data suggests that the functional consequences of mutations at this codon may be subtype-specific.

In our dataset, there are also instances where IN or RT mutations that were previously statistically associated with decreased RC when present in patient-derived sequences, were not

confirmed by mutagenesis experiments to significantly alter RC. A possibility is that these mutations are not acting alone to alter viral RC and may be linked to other mutations in the virus.

Conclusion

RT-integrase variants in or flanking CTL epitopes in the vital domains of the RT palm (158S), RT thumb (241I and 257V) and integrase CCD (124N) were confirmed to significantly reduce RC by mutagenesis experiments and represent potential vulnerable targets in Pol for an attenuation-based vaccine. Functional studies and structural models provide some insights into potential underlying mechanisms of how these mutations in RT may affect protein activity and amino acid interactions in the helix clamp, primer grip and catalytic site regions of RT.

Declaration of competing interest

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

Author contributions

DO: experiment design and execution, structural modelling, data analysis, writing of original draft. PM: qPCR data analysis, review & editing of the manuscript. MG: Supervision of modelling work, review & editing of the manuscript. TN: Conceptualization, funding acquisition, review & editing of manuscript, JM: Conceptualization, experiment design, data validation and investigation, review & editing of the manuscript, supervision and funding acquisition.

Acknowledgements

This work was funded by the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative

is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant # 107752/Z/15/Z] and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government. TN received additional funding from the South African Department of Science and Technology through the National Research Foundation (South African Research Chairs Initiative), the Victor Daitz Foundation and the Howard Hughes Medical Institute. This work was also partially supported by the Bill and Melinda Gates Foundation, the International AIDS Vaccine Initiative (IAVI) (UKZNRSA1001), the NIAID (R37AI067073) and Gilead Sciences (grant #00406). JKM received additional funding from the Poliomyelitis Research Foundation of South Africa and the South African Medical Research Council under a Self-Initiated Research Grant. DBAO received funding from National Research Foundation (NRF-SA) and the College of Health sciences, University of KwaZulu-Natal. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are grateful to all study participants and support staff.

References

1. Kiepiela, P., et al., *CD8+ T-cell responses to different HIV proteins have discordant associations with viral load.* Nature medicine, 2007. **13**(1): p. 46-53.
2. Koup, R., et al., *Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome.* Journal of virology, 1994. **68**(7): p. 4650-4655.
3. McMichael, A.J. and W.C. Koff, *Vaccines that stimulate T cell immunity to HIV-1: the next step.* Nature immunology, 2014. **15**(4): p. 319.
4. Migueles, S.A. and M. Connors, *Success and failure of the cellular immune response against HIV-1.* Nature Immunology, 2015. **16**(6): p. 563-570.

5. Schmitz, J.E., et al., *Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes.* Science, 1999. **283**(5403): p. 857-860.
6. Price, D.A., et al., *Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection.* Proceedings of the National Academy of Sciences, 1997. **94**(5): p. 1890-1895.
7. Allen, T.M. and M. Altfeld, *Crippling HIV one mutation at a time.* The Journal of experimental medicine, 2008. **205**(5): p. 1003-1007.
8. Chopera, D.R., et al., *Immune-mediated attenuation of HIV-1.* Future virology, 2011. **6**(8): p. 917-928.
9. Troyer, R.M., et al., *Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response.* PLoS Pathogens, 2009. **5**(4): p. e1000365.
10. Dahirel, V., et al., *Coordinate linkage of HIV evolution reveals regions of immunological vulnerability.* Proceedings of the National Academy of Sciences, 2011. **108**(28): p. 11530-11535.
11. Gaiha, G.D., et al., *Structural topology defines protective CD8+ T cell epitopes in the HIV proteome.* Science, 2019. **364**(6439): p. 480-484.
12. Rolland, M., et al., *HIV-1 conserved-element vaccines: relationship between sequence conservation and replicative capacity.* Journal of virology, 2013. **87**(10): p. 5461-5467.
13. Liu, Y., et al., *Impact of mutations in highly conserved amino acids of the HIV-1 Gag-p24 and Env-gp120 proteins on viral replication in different genetic backgrounds.* PLoS One, 2014. **9**(4): p. e94240.
14. Ferguson, A.L., et al., *Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design.* Immunity, 2013. **38**(3): p. 606-617.
15. Mann, J.K., et al., *The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing.* PLoS computational biology, 2014. **10**(8): p. e1003776.
16. Crawford, H., et al., *Evolution of HLA-B* 5703 HIV-1 escape mutations in HLA-B* 5703–positive individuals and their transmission recipients.* The Journal of experimental medicine, 2009. **206**(4): p. 909-921.
17. Wright, J.K., et al., *Impact of HLA-B* 81-associated mutations in HIV-1 Gag on viral replication capacity.* Journal of virology, 2012. **86**(6): p. 3193-3199.
18. Shahid, A., et al., *Consequences of HLA-B*13-Associated Escape Mutations on HIV-1 Replication and Nef Function.* Journal of Virology, 2015. **89**(22): p. 11557-11571.
19. Domingo, E., *Variation of the HIV Genome: Implications for the Pathogenesis and Prevention of AIDS*, in *RNA Genetics*. 2018, CRC Press. p. 157-168.
20. Borthwick, N., et al., *Vaccine-elicited human T cells recognizing conserved protein regions inhibit HIV-1.* Molecular Therapy, 2014. **22**(2): p. 464-475.
21. Gandhi, R.T., et al., *Progressive reversion of human immunodeficiency virus type 1 resistance mutations in vivo after transmission of a multiply drug-resistant virus.* Clinical infectious diseases, 2003. **37**(12): p. 1693-1698.
22. Mothe, B., et al., *A human immune data-informed vaccine concept elicits strong and broad T-cell specificities associated with HIV-1 control in mice and macaques.* Journal of translational medicine, 2015. **13**(1): p. 60.
23. Mothe, B., et al., *Definition of the viral targets of protective HIV-1-specific T cell responses.* Journal of translational medicine, 2011. **9**(1): p. 208.
24. Hu, Z. and D.R. Kuritzkes, *Effect of raltegravir resistance mutations in HIV-1 integrase on viral fitness.* J Acquir Immune Defic Syndr, 2010. **55(2)**: p. 148-55. doi: 10.1097/QAI.0b013e3181e9a87a.
25. Weber, J., et al., *Diminished replicative fitness of primary human immunodeficiency virus type 1 isolates harboring the K65R mutation.* J Clin Microbiol, 2005. **43(3)**: p. 1395-1400.

26. Gandhi, R.T., et al., *Progressive reversion of human immunodeficiency virus type 1 resistance mutations in vivo after transmission of a multiply drug-resistant virus.* Clin Infect Dis, 2003. **37(12)**: p. 1693-1698.
27. Theys, K., et al., *Within-patient mutation frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in HIV.* PLoS genetics, 2018. **14**(6): p. e1007420.
28. Brumme, Z.L., et al., *Reduced replication capacity of NL4-3 recombinant viruses encoding reverse transcriptase-integrase sequences from HIV-1 elite controllers.* J Acquir Immune Defic Syndr, 2011. **56(2)**: p. 100-8. doi: 10.1097/QAI.0b013e3181fe9450.
29. Brockman, M.A., et al., *Uncommon pathways of immune escape attenuate HIV-1 integrase replication capacity.* J Virol, 2012. **86(12)**: p. 6913-6923. doi: 10.1128/JVI.07133-11.
30. Rihn, S.J., et al., *Uneven genetic robustness of HIV-1 integrase.* J Virol, 2015. **89**(1): p. 552-67.
31. Ojwach, D.B.A., et al., *"Pol-Driven Replicative Capacity Impacts Disease Progression in HIV-1 Subtype C Infection".* J Virol, 2018. **93**(3).
32. Brockman, M.A., et al., *Uncommon pathways of immune escape attenuate HIV-1 integrase replication capacity.* Journal of virology, 2012. **86**(12): p. 6913-6923.
33. Demeulemeester, J., et al., *HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort.* Cell Host Microbe, 2014. **16**(5): p. 651-62.
34. Frater, A.J., et al., *Effective T-cell responses select human immunodeficiency virus mutants and slow disease progression.* Journal of virology, 2007. **81**(12): p. 6742-6751.
35. Kløverpris, H.N., et al., *CD8+ TCR bias and immunodominance in HIV-1 infection.* The Journal of Immunology, 2015: p. 1400854.
36. Oxenius, A., et al., *Loss of viral control in early HIV-1 infection is temporally associated with sequential escape from CD8+ T cell responses and decrease in HIV-1-specific CD4+ and CD8+ T cell frequencies.* Journal of Infectious Diseases, 2004. **190**(4): p. 713-721.
37. Allen, T.M., et al., *Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution.* J Virol, 2005. **79**(21): p. 13239-49.
38. Los Alamos National Laboratory. *HIV Molecular Database.* 2020 [cited 2020 1/05/2020]; Available from: http://www.hiv.lanl.gov
39. Miura, T., et al., *HLA-associated alterations in replication capacity of chimeric NL4-3 viruses carrying gag-protease from elite controllers of human immunodeficiency virus type 1.* Journal of virology, 2009. **83**(1): p. 140-149.
40. Brumme, Z.L., et al., *Reduced replication capacity of NL4-3 recombinant viruses encoding RT-integrase sequences from HIV-1 elite controllers.* Journal of acquired immune deficiency syndromes (1999), 2011. **56**(2).
41. Gervaix, A., et al., *A new reporter cell line to monitor HIV infection and drug susceptibility in vitro.* Proceedings of the National Academy of Sciences, 1997. **94**(9): p. 4653-4658.
42. Brockman, M.A., et al., *Use of a novel GFP reporter cell line to examine replication capacity of CXCR4-and CCR5-tropic HIV-1 by flow cytometry.* Journal of virological methods, 2006. **131**(2): p. 134-142.
43. Van Maele, B., et al., *Impact of the central polypurine tract on the kinetics of human immunodeficiency virus type 1 vector transduction.* Journal of virology, 2003. **77**(8): p. 4685-4694.
44. Butler, S.L., M.S.T. Hansen, and F.D. Bushman, *A quantitative assay for HIV DNA integration in vivo.* Nature Medicine, 2001. **7**(5): p. 631-634.
45. De Houwer, S., et al., *The HIV-1 integrase mutant R263A/K264A is 2-fold defective for TRN-SR2 binding and viral nuclear import.* Journal of Biological Chemistry, 2014. **289**(36): p. 25351-25361.

46. Schmittgen, T.D. and B.A. Zakrajsek, *Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR.* J Biochem Biophys Methods, 2000. **46**(1-2): p. 69-81.
47. Waterhouse, A., et al., *SWISS-MODEL: homology modelling of protein structures and complexes.* Nucleic Acids Res, 2018. **46**(W1): p. W296-w303.
48. Pettersen, E.F., et al., *UCSF Chimera—a visualization system for exploratory research and analysis.* Journal of computational chemistry, 2004. **25**(13): p. 1605-1612.
49. Sharma, B., et al., *A positively charged side chain at position 154 on the β8–αE loop of HIV-1 RT is required for stable ternary complex formation.* Nucleic Acids Research, 2003. **31**(17): p. 5167-5174.
50. Sharaf, N.G., et al., *NMR structure of the HIV-1 reverse transcriptase thumb subdomain.* Journal of biomolecular NMR, 2016. **66**(4): p. 273-280.
51. Hermann, T., et al., *The 'helix clamp'in HIV-1 reverse transcrptase: a new nucleic acide binding motif common in nucleic acid polymerases.* Nucleic acids research, 1994. **22**(22): p. 4625-4633.
52. Beard, W.A., et al., *Structure/function studies of human immunodeficiency virus type 1 reverse transcriptase. Alanine scanning mutagenesis of an alpha-helix in the thumb subdomain.* Journal of Biological Chemistry, 1994. **269**(45): p. 28091-28097.
53. Bebenek, K., et al., *Reduced frameshift fidelity and processivity of HIV-1 reverse transcriptase mutants containing alanine substitutions in helix H of the thumb subdomain.* Journal of Biological Chemistry, 1995. **270**(33): p. 19516-19523.
54. Dunn, L.L., et al., *Mutations in the Thumb Allow Human Immunodeficiency Virus Type 1 Reverse Transcriptase To Be Cleaved by Protease in Virions.* Journal of Virology, 2009. **83**(23): p. 12336-12344.
55. Palaniappan, C., et al., *Mutations within the primer grip region of HIV-1 reverse transcriptase result in loss of RNase H function.* Journal of Biological Chemistry, 1997. **272**(17): p. 11157-11164.
56. Yu, Q., et al., *Mutations in the primer grip of human immunodeficiency virus type 1 reverse transcriptase impair proviral DNA synthesis and virion maturation.* Journal of virology, 1998. **72**(9): p. 7676-7680.
57. Wrobel, J.A., et al., *A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase.* Proceedings of the National Academy of Sciences, 1998. **95**(2): p. 638-645.
58. Herzig, E., et al., *A Novel Leu92 Mutant of HIV-1 Reverse Transcriptase with a Selective Deficiency in Strand Transfer Causes a Loss of Viral Replication.* Journal of virology, 2015. **89**(16): p. 8119-8129.
59. Bartha, I., et al., *A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control.* Elife, 2013. **2**: p. e01123.
60. Kessl, J., et al., *HIV-1 Integrase-DNA Recognition Mechanisms.* Viruses, 2009. **1**(3): p. 713.
61. Zhao, Z., et al., *Subunit-specific protein footprinting reveals significant structural rearrangements and a role for N-terminal Lys-14 of HIV-1 Integrase during viral DNA binding.* Journal of Biological Chemistry, 2008. **283**(9): p. 5632-5641.
62. Gupta, K., et al., *Structural basis for inhibitor-induced aggregation of HIV integrase.* PLoS biology, 2016. **14**(12).
63. Brenner, B.G., et al., *Subtype diversity associated with the development of HIV-1 resistance to integrase inhibitors.* Journal of medical virology, 2011. **83**(5): p. 751-759.

Vulnerable targets in HIV-1 Pol for attenuation-based vaccine design

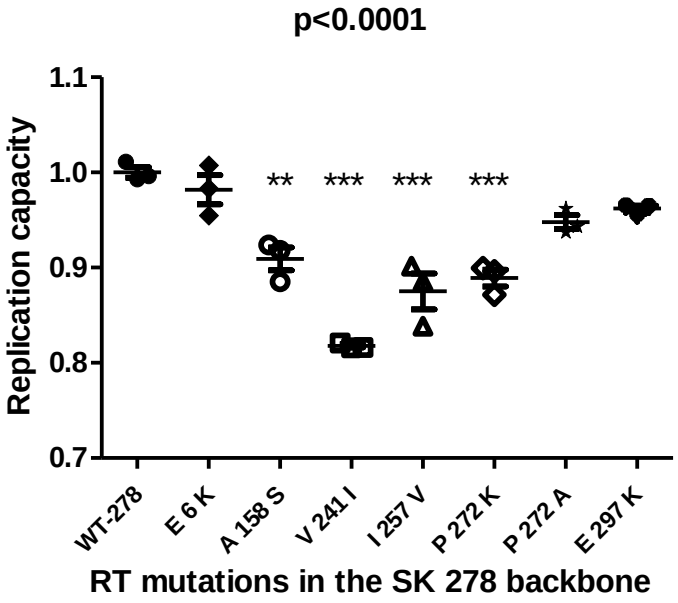


Figure 1: Replication capacities of NL4-3 recombinant viruses encoding wildtype (WT) SK-278 RT-integrase and SK-278 with RT mutations

The RT mutations A158S, V241I, I257V and P272K significantly reduced RC expressed relative to the wildtype. Bars represent the means for at least 3 independent experiments, and error bars represent standard deviations from the means. ANOVA with Tukey post hoc tests was used to test for significance. The ANOVA p value is shown and the asterisks denote significant post hoc tests.

*-is equivalent to $p < 0.05$; ** to $p < 0.01$ and *** to $p < 0.001$, RT-reverse transcriptase, RC-replication capacity

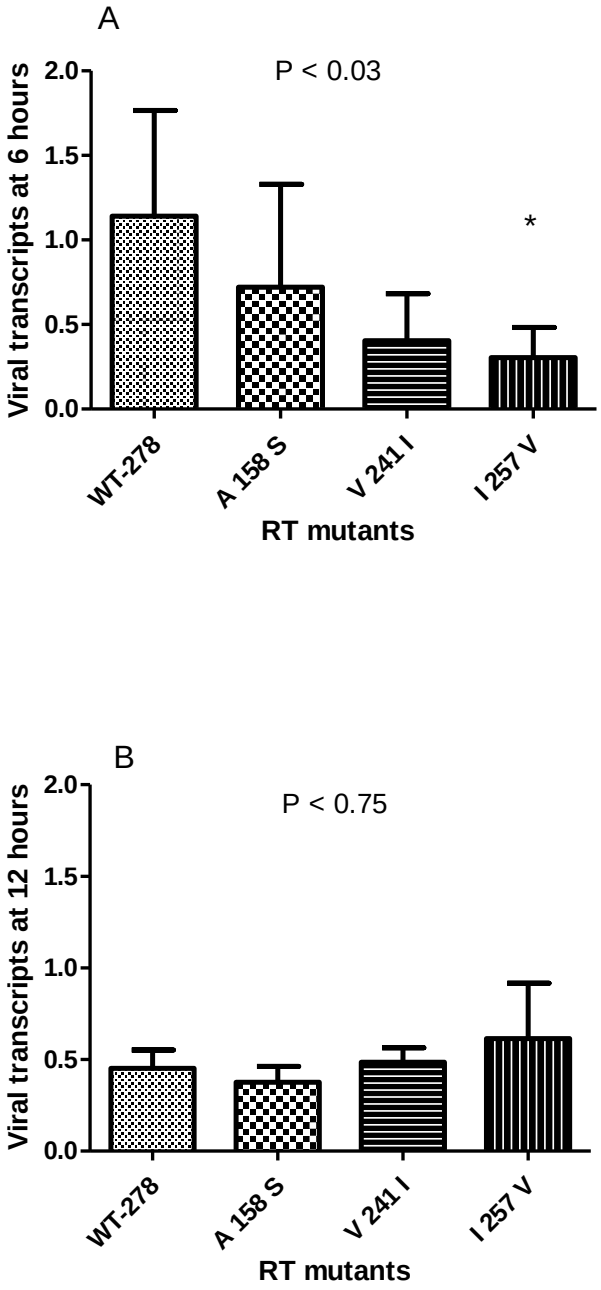


Figure 2: Reverse transcription of NL4-3 recombinant viruses encoding wildtype (WT) SK-278 RT-integrase and SK-278 with RT mutations

The viral transcripts measured from the RT mutant viruses harbouring the A158S, V241I and I257V variants were reduced compared to the wildtype at 6 hours (**A**) but not 12 hours (**B**) post-infection expressed relative to the wildtype. Bars represent the means for duplicate measurements from 3 independent infections, and error bars represent standard deviations from the means. ANOVA with Tukey post hoc tests was used to test for significance. The ANOVA p value is shown and the asterisks denote significant post hoc tests.

*-is equivalent to $p < 0.05$; ** to $p < 0.01$ and *** to $p < 0.001$, RT-reverse transcriptase, RC-replication capacity

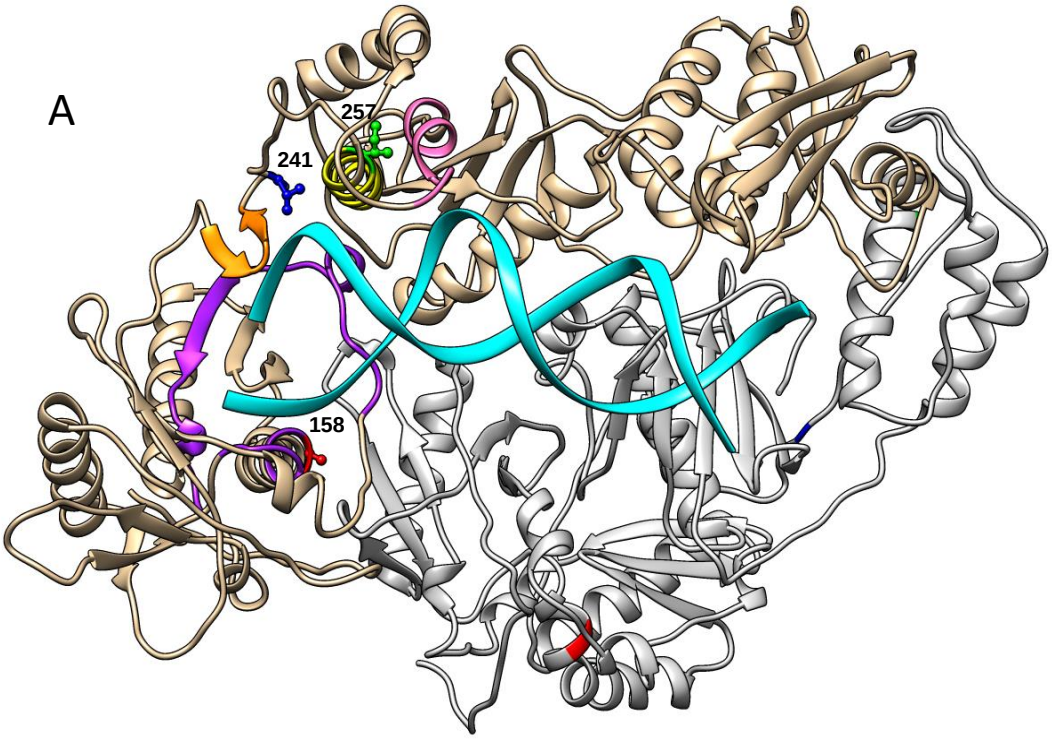
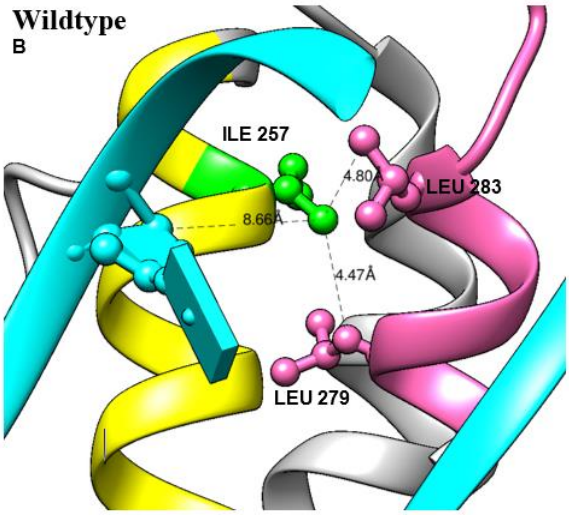
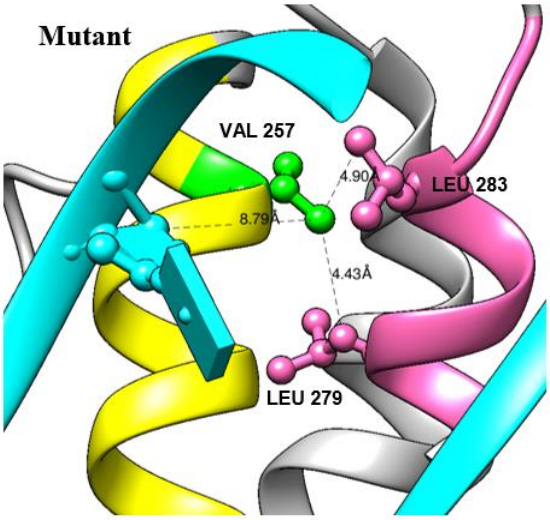
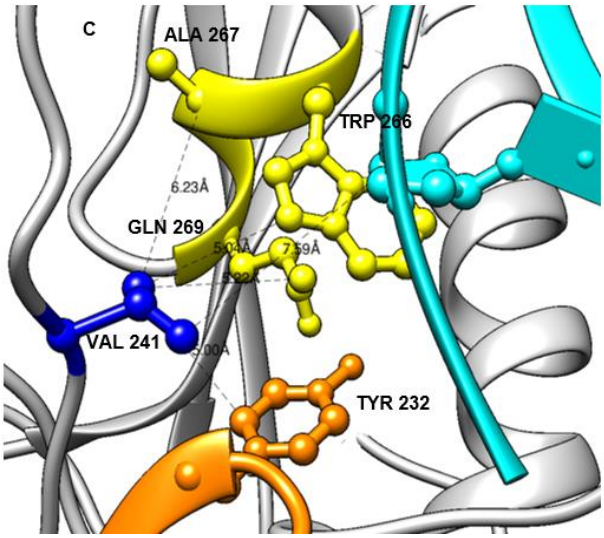


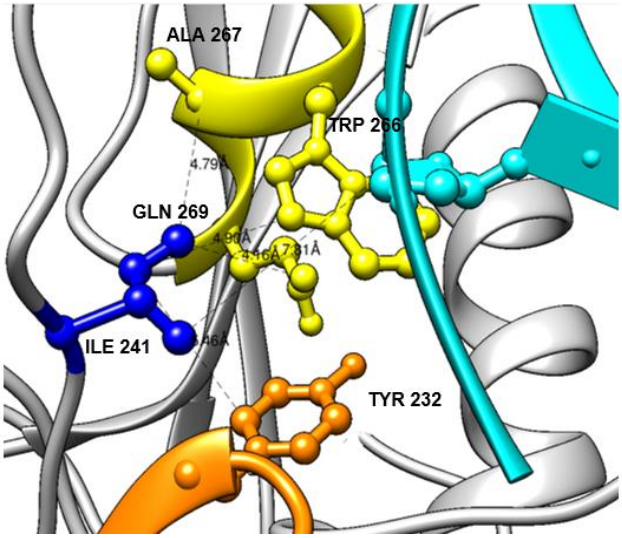
Figure 3: RT models of the SK 278 wildtype sequence and mutants

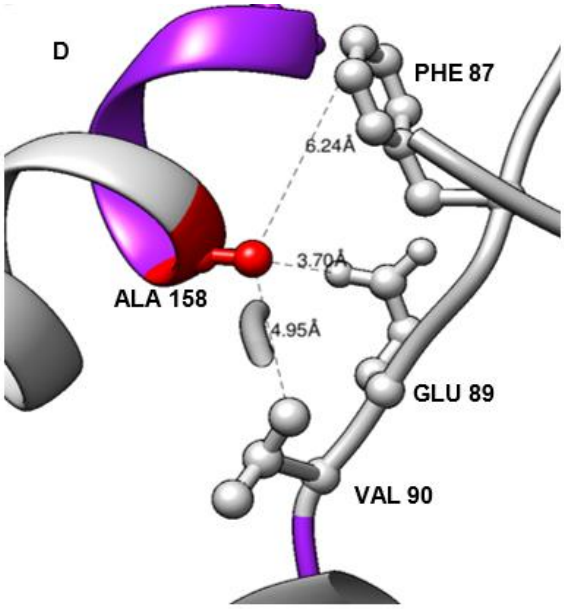
A. A ribbon diagram of the of the HIV-1 RT SK 278 sequence is shown, where the p66 subunit (brown) and p51 subunit (grey) are in complex with nucleic acid template (cyan). The alpha-H (yellow) and alpha-I (pink) helices as well as the primer grip (orange) of the thumb region are highlighted. The catalytic site is shown in purple. The mutant residues reported are at codons 158 (red), 241(blue) and 257 (green) with their atoms labelled in ball-and-stick format. **B.** The interaction of codon 257, for both wildtype and mutant residues, with the alpha-I helix and the template is shown. **C.** The interaction of codon 241, for both wildtype and mutant residues, with the alpha-H helix and primer grip is shown. **D.** The interaction of codon 241, for both wildtype and mutant residues, with the β5a strand is shown. The mutant forms a h-bond (red) with codon 155. The distances marked in panels B-D are where the maximum differences between wildtype and mutant models were observed. Models were performed using SWISSMODEL and the molecular graphics and analyses were performed with UCSF Chimera (2).

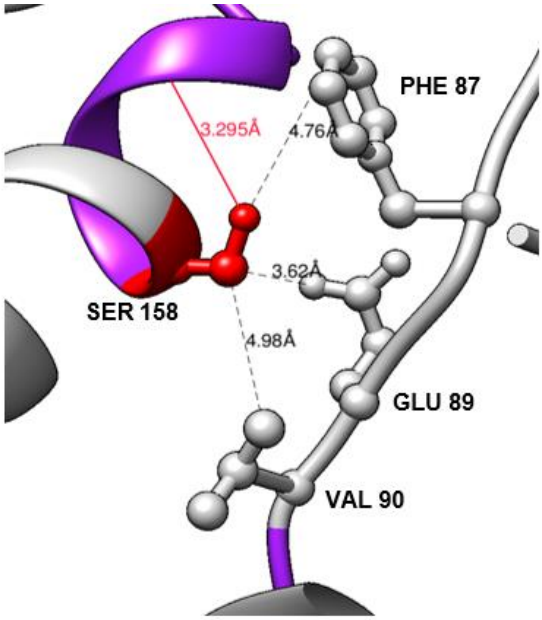

Wildtype
B
ILE 257
LEU 283
LEU 279
8.66Å
4.80Å
4.47Å


Mutant
VAL 257
LEU 283
LEU 279
8.79Å
4.90Å
4.43Å


C
ALA 267
TRP 266
GLN 269
VAL 241
TYR 232
6.23Å
7.59Å


ALA 267
TRP 266
GLN 269
ILE 241
TYR 232


D
PHE 87
ALA 158
GLU 89
VAL 90
6.24Å
3.70Å
4.95Å


PHE 87
SER 158
GLU 89
VAL 90
3.295Å
4.76Å
3.62Å
4.98Å

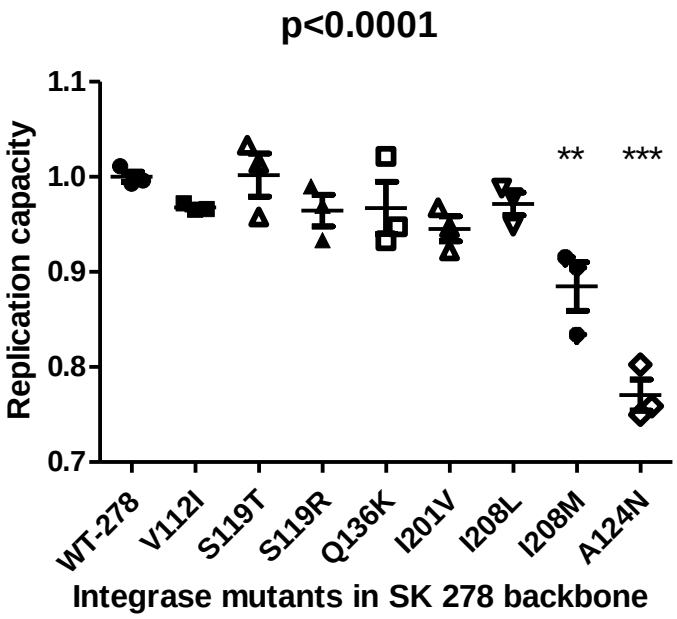


Figure 4: Replication capacity of NL4-3 recombinant viruses encoding wild-type SK-278 RT-integrase and SK-278 with integrase mutations

The effect of various integrase mutations in the SK-278 RT-integrase is shown. The mutations 124N and 208M significantly reduced RC relative to the SK-278 wild-type. Bars represent the means for at least 3 independent experiments, and error bars represent standard deviations from the means. ANOVA with Tukey post hoc tests was used to test for significance. The ANOVA p value is shown and the asterisks denote significant post hoc tests.

*-is equivalent to $p < 0.05$; ** to $p < 0.01$ and *** to $p < 0.001$, SK-Sinikithemba, ANOVA-analysis of variance

Supplementary Table 1: Titre results and data for triplicate independent replication capacity (RC) assays are shown for viruses encoding mutations in reverse transcriptase (RT) and integrase (IN).

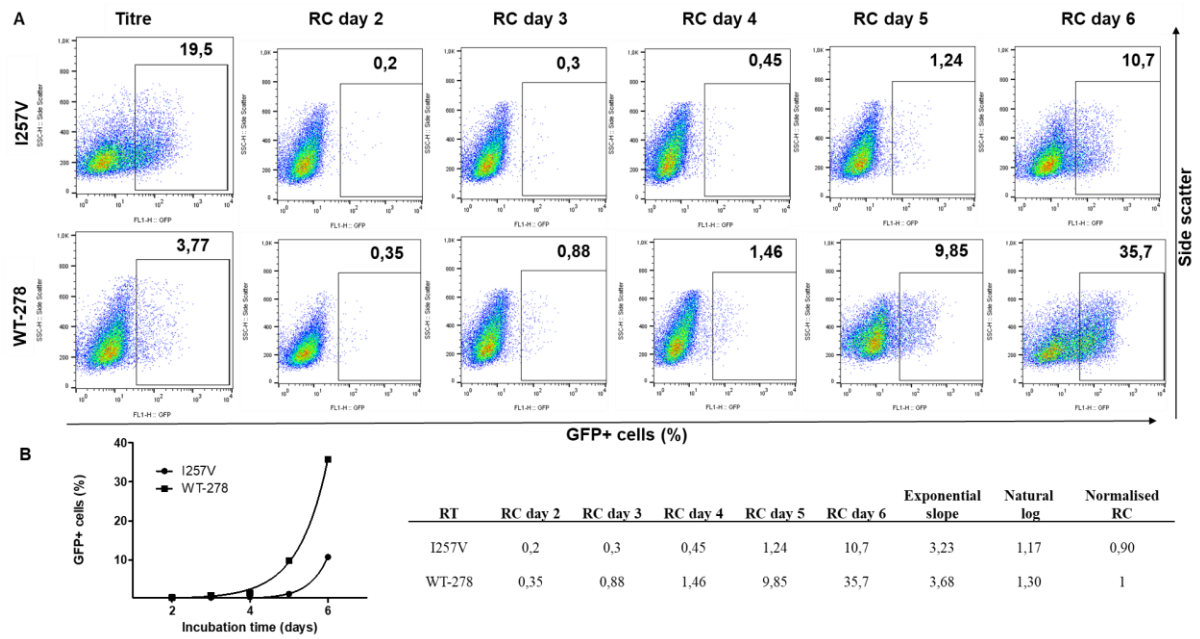
RT mutant	Titre (GFP)[a]	Virus (µl)[b]	Media (µl)[c]	RC1[d]	RC2[d]	RC3[d]
E 6 K	5,87%	20,44	379,56	1,01	0,95	0,98
A 158 S	4,29%	27,97	372,03	0,92	0,89	0,92
V 241 I	2,55%	47,06	352,94	0,82	0,82	0,82
I 257 V	19,50%	6,15	393,85	0,90	0,84	0,89
P 272 K	6,24%	19,23	380,77	0,90	0,90	0,87
E 297 K	3,13%	38,34	361,66	0,97	0,96	0,97
P 272 A	5,82%	20,62	379,38	0,94	0,96	0,94
WT-278	3,77%	31,83	368,17	0,99	1	1,01
IN mutant						
V112I	11,00%	10,91	389,09	0,97	0,97	0,97
S119R	5,43%	22,1	377,9	0,93	0,99	0,97
Q136K	2,80%	42,86	357,14	1,02	0,95	0,93
I201V	8,23%	14,58	385,42	0,97	0,92	0,95
I208L	5,38%	22,3	377,7	0,95	0,98	0,99
I208M	5,45%	22,02	377,98	0,90	0,83	0,92
A124N	4,63%	25,92	374,08	0,80	0,76	0,75
S119T	3,04%	39,47	360,53	1,03	1,01	0,96
WT-278	3,77%	31,83	368,17	0,99	1	1,01

a The percentage of green fluorescent protein (GFP) expressing cells obtained in the titration assay following infection of 1 million cells with 400 µl virus.

b The volume of virus used to achieve 0.3 % infection in the RC assay as calculated from the titre result.

c The virus volume used for infection in the RC assay was made up to a final volume of 400 µl using media.

d The values for the natural log of the slope of increase in GFP positive cells for the mutants are divided by that of the wildtype to express the RC of the mutants relative to the wildtype (RC = 1).



RT	RC day 2	RC day 3	RC day 4	RC day 5	RC day 6	Exponential slope	Natural log	Normalised RC
I257V	0,2	0,3	0,45	1,24	10,7	3,23	1,17	0,90
WT-278	0,35	0,88	1,46	9,85	35,7	3,68	1,30	1

Supplementary Figure S1. Flow cytometry plots showing the spread of infection of the I257V mutant and WT-278 and calculation of replication capacity (RC).

Panel A. Flow cytometry plots obtained for the titre and RC assay for the wildtype virus (WT-278) and the I257V mutant virus are shown. The percentage GFP (green fluorescent protein) cells obtained in the titre was used to normalize the starting percentage infection to 0.3 % for the RC assay. The increase in the % of GFP positive cells from days 2 to 6 is shown for both the wildtype and mutant viruses. Panel B. The increase in % GFP cells over time in the RC assay is plotted and the calculation of the RC value of each virus from the slope of this exponential spread of infection is shown.

Supplementary Table 2: Distance between residues of interest (I257V, V241I and A158S) and interacting residues/templates, in wildtype and mutant models.

Residue of mutation	Interacting residue/template	Wildtype model	Mutant model	Difference in distance[a]	Figure
I257V	Leu 279	4.47	4.43	0.04	3B
	Leu 283	4.80	4.90	-0.10	
	DNA template	8.66	8.79	-0.13	
V241I	Trp 266	5.04	4.90	0.14	3C
	Ala 267	6.23	4.79	1.44	
	Gln 269	5.22	4.16	1.06	
	Tyr 232	5.00	5.46	-0.46	
	DNA Template	7.59	7.81	-0.21	
A158S	Val 90	4.95	4.98	-0.03	3D
	Glu 89	3.70	3.62	0.08	
	Phe 87	6.24	4.76	1.48	
	H-bond at 155	No	Yes at 3.295		

[a] The difference between the wildtype and mutant models for the distance between the two residues indicated.