Computational Design of gRNAs Targeting Genetic Variants Across HIV-1 Subtypes for CRISPR-Mediated Antiviral Therapy

Cheng-Han Chung[1,2], Alexander G. Allen[1,2], Andrew Atkins[1,2], Robert W. Link[1,2], Michael R. Nonnemacher[1,2,3], Will Dampier[1,2] and Brian Wigdahl[1,2,3]*

[1] Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, PA, United States, [2] Center for Molecular Virology and Translational Neuroscience, Institute for Molecular Medicine and Infectious Disease, Drexel University College of Medicine, Philadelphia, PA, United States, [3] Sidney Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, PA, United States

Clustered regularly interspaced short palindromic repeats (CRISPR)-based HIV-1 genome editing has shown promising outcomes in *in vitro* and *in vivo* viral infection models. However, existing HIV-1 sequence variants have been shown to reduce CRISPR-mediated efficiency and induce viral escape. Two metrics, global patient coverage and global subtype coverage, were used to identify guide RNA (gRNA) sequences that account for this viral diversity from the perspectives of cross-patient and cross-subtype gRNA design, respectively. Computational evaluation using these parameters and over 3.6 million possible 20-bp sequences resulted in nine lead gRNAs, two of which were previously published. This analysis revealed the benefit and necessity of considering all sequence variants for gRNA design. Of the other seven identified novel gRNAs, two were of note as they targeted interesting functional regions. One was a gRNA predicted to induce structural disruption in the nucleocapsid binding site (Ψ), which holds the potential to stop HIV-1 replication during the viral genome packaging process. The other was a reverse transcriptase (RT)-targeting gRNA that was predicted to cleave the subdomain responsible for dNTP incorporation. CRISPR-mediated sequence edits were predicted to occur on critical residues where HIV-1 has been shown to develop resistance against antiretroviral therapy (ART), which may provide additional evolutionary pressure at the DNA level. Given these observations, consideration of broad-spectrum gRNAs and cross-subtype diversity for gRNA design is not only required for the development of generalizable CRISPR-based HIV-1 therapy, but also helps identify optimal target sites.

Keywords: human immunodeficiency virus type 1 (HIV-1), CRISPR/Cas9, gene therapy, genetic variation, HIV-1 subtypes, bioinformatics, gRNA design

INTRODUCTION

Human immunodeficiency virus type 1 (HIV-1) has been recognized as the causative agent of acquired immunodeficiency syndrome (AIDS) since 1983. Despite controversial evidence for the identification of viral strains that caused zoonotic transmission, phylogenetic analyses suggest that HIV-1 originated from cross-species transmission of simian immunodeficiency virus (SIV) from non-human primates to humans (Keele et al., 2006; Van Heuverswyn et al., 2007; de Silva et al., 2008). Independent zoonotic transmissions of HIV resulted in distinct lineages of HIV-1 viruses that are termed the M, N, O, and P groups (De Leys et al., 1990; Simon et al., 1998; Roques et al., 2004; Vallari et al., 2011). Group M represents the global HIV-1 pandemic with approximately 38 million people living with HIV-1 (Fauci, 2017; World Health Organization, 2020). Due to the rapid genetic divergence during HIV-1 replication coupled with geographic constraints, distinctive lineages within the group M phylogeny evolved independently. These phylogenetic observations were used to designate a collection of HIV-1 subtypes (Hemelaar, 2012). Currently, the field recognizes nine major subtypes (A, B, C, D, F, G, H, J, and K), 96 distinct circulating recombinant forms (CRFs), and unique recombinant forms (URFs) that lack prevalent transmission (Robertson et al., 2000). The average genetic distance within subtypes falls between 8% and 17% with outliers as high as 30% (Korber et al., 2001). Genetic diversity between subtypes ranges from 17% to 42% and has likely been increasing due to evolving recombinant forms (Rambaut et al., 2001; Abecasis et al., 2009).

HIV-1 sequence variation originates from a lack of proofreading during HIV-1 reverse transcription within infected individuals (Hu and Hughes, 2012). Selection pressure due to host immune responses including cell killing and HIV-1-specific antibodies further impact the level of sequence diversity (Phillips et al., 1991; Wei et al., 2003). Error-prone replication and rapid adaptation to host immunity resulted in higher variation within HIV-1 structural genes while viral enzymes diversified at a lower rate (Yu et al., 2004; Kearney et al., 2009; Zanini et al., 2015). Antiretroviral therapy (ART) effectively reduces plasma viral load to undetectable levels, which largely decreases host diversity within a person (Maldarelli et al., 2007; Kearney et al., 2014). However, the establishment of latency across different cellular and anatomical compartments has made ART insufficient to cure HIV-1 infection (Sturdevant et al., 2015; Barton et al., 2016; Bui et al., 2017; Hosmane et al., 2017). Low levels of viral replication within cellular and anatomical reservoirs have also been shown to increase viral diversity under suppressive ART (Palmer et al., 2008; Dampier et al., 2014; Dampier et al., 2016).

Viral rebound has been found in most clinical studies after ART cessation (Persaud and Luzuriaga, 2014; Shah et al., 2014; Henrich et al., 2017; Colby et al., 2018). The major hurdle for developing an HIV-1 cure has been the integrated provirus within the latent reservoir that evades host immune surveillance and ART. Genome editing using the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 system has recently been applied to inactivate HIV-1 replication

or excise proviral DNA (Ebina et al., 2013; Hu et al., 2014; Zhu et al., 2015; Yin et al., 2016; Kaminski et al., 2016a; Wang et al., 2016a; Kaminski et al., 2016b; Wang et al., 2016b; Lebbink et al., 2017; Yin et al., 2017; Zhao et al., 2017; Bella et al., 2018; Ophinni et al., 2018; Wang Q. et al., 2018; Wang Z. et al., 2018; Dampier et al., 2018a; Dash et al., 2019; Sullivan et al., 2019; Chung C-H et al., 2020). CRISPR-Cas9 is comprised of a guide RNA (gRNA), a 20-nt RNA sequence designed to specify the genomic target that is contained on a larger structural RNA molecule, and the Cas9 endonuclease. This ribonucleoprotein complex facilitates target-specific gene editing using the gRNA to pair with target DNA (also termed as the protospacer), following the recognition of a protospacer adjacent motif (PAM) (Cong et al., 2013; Jinek et al., 2013). Upon sufficient recognition, Cas9 induces double-strand breaks (DSBs) within the target DNA. DSBs will then facilitate endogenous DNA repair mainly via non-homologous end joining (NHEJ) in the absence of donor template. The error-prone process of NHEJ-mediated repair has been shown to introduce insertions/deletions (InDels) that render deleterious effects on target genes.

The gene targeting mechanism of CRISPR mediated by the gRNA sequence and the presence of a PAM site has created the opportunity to edit proviral DNA at the site of interest based on functional domains. The 20-bp gRNA sequence that matches the 20-bps of HIV-1 DNA adjacent to a PAM site has dictated the editing region within the HIV-1 provirus. However, the effects of CRISPR-mediated editing vary due to different selection of HIV-1-targeting sites. CRISPR-mediated HIV-1 inactivation could be ineffective by the selection of target sites that are tolerant to InDels without affecting HIV-1 replication. This variation of CRISPR-mediated HIV-1 inactivation has been found in multiple CRISPR screening studies using gRNAs targeting viral genes spanning the HIV-1 proviral genome (Yin et al., 2016; Wang et al., 2016b). HIV-1 diversity further poses critical challenges to the development of this antiviral strategy. Previous screening studies have shown that certain gRNA-protospacer mismatches reduce CRISPR-mediated editing efficiency (Hsu et al., 2013; Doench et al., 2016). This has indicated that even before CRISPR-mediated editing begins, the presence of mismatches between gRNA sequences and HIV-1 target sites reduce the overall editing efficiency within the existing reservoir. Previous studies have shown that patient-derived molecular clones that contained one or more gRNA-protospacer mismatches within the gag and env gene resulted in escape mutants during long-term culture up to 110 days (Darcis et al., 2019).

Editing efficiency is reduced when mismatches exist between gRNAs and target sequences (Hsu et al., 2013; Doench et al., 2016). However, this has been found to be non-linear across the 20-bp protospacer. Position-specific penalty matrices have been developed to quantify the editing efficiency with the presence of gRNA-target mismatches (Hsu et al., 2013; Doench et al., 2016). Hsu et al. developed the MIT matrix, a 20-number vector to indicate a position-specific penalty regardless of mismatch type (Hsu et al., 2013). The cutting frequency determination (CFD) table, a 16x20 matrix, was later developed to incorporate both position and nucleotide specific mismatch-when assigning an

editing efficiency penalty (Doench et al., 2016). We and others have demonstrated that the CFD matrix gives better editing-efficiency predictions than the MIT matrix (Cui et al., 2018; Chung C.-H. et al., 2020). Given this observation, the CFD matrix was selected to estimate editing efficiency between candidate gRNAs and HIV-1 genetic variants. In general, a mismatch at PAM-distal regions was more tolerable than PAM-proximal mismatches. In the context of HIV-1 sequence variants, a candidate gRNA may still induce CRISPR-mediated DSBs at the intended target sites if the mismatches existed at PAM-distal side. The CFD matrix assigns a score between 0 and 1, where 1 represents optimal editing efficiency. We have identified a CFD cutoff score of 0.569 to binarize the chance of inducing DSB using publicly available datasets (Haeussler et al., 2016; Chung C.-H. et al., 2020). A CFD score above 0.569 was predicted to induce target editing 95% of the time (Chung C.-H. et al., 2020).

The overall objective of the present study was to propose optimal target sites across HIV-1 subtypes, as well as computationally identify optimal gRNA sequences within selected target sites that cover most existing genetic variants. These target sites will be ideal candidates for developing a pool of gRNAs for targeting any HIV-1-infected individual as opposed to requiring an individually personalized approach. We hypothesize a truly personalized approach would be resource limiting by the need for safety and efficacy trials for each new construct. To accomplish this, we utilized an exhaustive gRNA search pipeline previously developed using k-mer analysis and a position-specific penalty matrix that describes the penalty for gRNA-target mismatches at different positions (Doench et al., 2016; Dampier et al., 2018a; Sullivan et al., 2019). Using this approach, target sites that were conserved across all major subtypes and CRFs were identified. These regions of low diversity also allowed high coverage of known variants in targeted sites across HIV-1 subtypes with carefully chosen gRNA sequences. These analyses resulted in a list of gRNAs proposed for generalized use across patients infected by different subtypes with high predicted CRISPR-mediated efficiency at HIV-1 replication inactivation and a low predicted occurrence of CRISPR resistance.

METHODS

HIV-1 Sequence Collection From the LANL Database

HIV-1 sequences were retrieved using the sequence search interface on the LANL HIV-1 sequence database with a specified organism of "HIV-1" in the sequence sample (SSAM) section without other limits to the rest of parameters (as of August 2018). Subtype of sequences was specified in the column designated "Subtype" of SSAM section in the LANL HIV01 sequence database. Minor subtypes, CRFs, URFs, and untyped samples were all categorized into 'Others'. All sequences were used for subsequent analysis as long as the tested gRNA was fully covered by the HIV-1 sequences across a 20-bp nucleotide stretch followed by the presence of a 3-bp PAM.

Calculation of Global Subtype Coverage, Global Patient Coverage, and Potential Off-Target Sites in the Human Genome

The Subtype Coverage (SC) of a designated gRNA was defined as the number of HIV-1 sequences that resulted in a CFD score above 0.569 divided by the total number of observed sequences within a subtype. The SC was calculated assuming that each sequence is an independent sample. The Patient Coverage (PC) was defined as the number of HIV-1 sequences that resulted in a CFD score above 0.569 divided by the total number of tested sequences observed in the same patient. The potential off-target site in the human genome was defined by any 20-bp genomic sequences that resulted in a CFD score above 0.569 compared with the 20-bp sequence of a tested gRNA. The Pat_SSID column in the LANL database was used to assign sequences to a single patient. The estimated global prevalence was used to weigh the effect of each subtype or patient in the global coverages.

Calculation of Sequence Diversity

The sequence diversity in this study was determined by a summation of Shannon entropy over a 20-bp window. The sequence diversity was defined as:

$$-\sum_{i=1}^{20}\sum_{j} p_{i,j} \times \log_2 p_{i,j}$$

where i is the nucleotide position within the 20-bp window, j is the nucleotide identity (A, C, G, T), and p is the nucleotide probability (e.g. $p_{10,A}$ represents the probability of A at the 10th bp in a given 20-bp window). Sequence diversity ranges between 0 and 40 bits. A sequence diversity of 40 bits means that every base at every position within the window is random. Conversely, a sequence diversity of 0 bits means that every sequence within this window has converged to one variant. The subtype diversity was calculated separately. The global diversity was calculated by the summation of subtype diversity weighted by the estimated global prevalence (Hemelaar et al., 2019).

Nomenclature of gRNA Identifier

To standardize gRNA names, the 20-bp spacer and corresponding PAM sequence were combined to form a single word used to generate an identifier for each unique gRNA. This conversion was performed by passing the unique gRNA word into the md5 hash function. This generated a 32-letter hexadecimal output. The first six letters from the output were used to generate that corresponding gRNA's identifier. Within the current number of published gRNAs, the chance of identifier collision should be relatively low.

Statistical and Bioinformatic Analysis

All bioinformatic analysis was conducted in Python. Simple linear regression analyses were conducted using the scipy.stats package (version=1.4.1). Root-mean-square error was calculated using the sklearn package (version=0.22). Alpha levels were set at 5% for the Wald Tests for the significance of β coefficients (slope) from simple linear regression models. The CRSeek package was used to calculate percent variant coverage and potential off-target cleavage sites adapting cas-offinder (Dampier et al., 2018b).
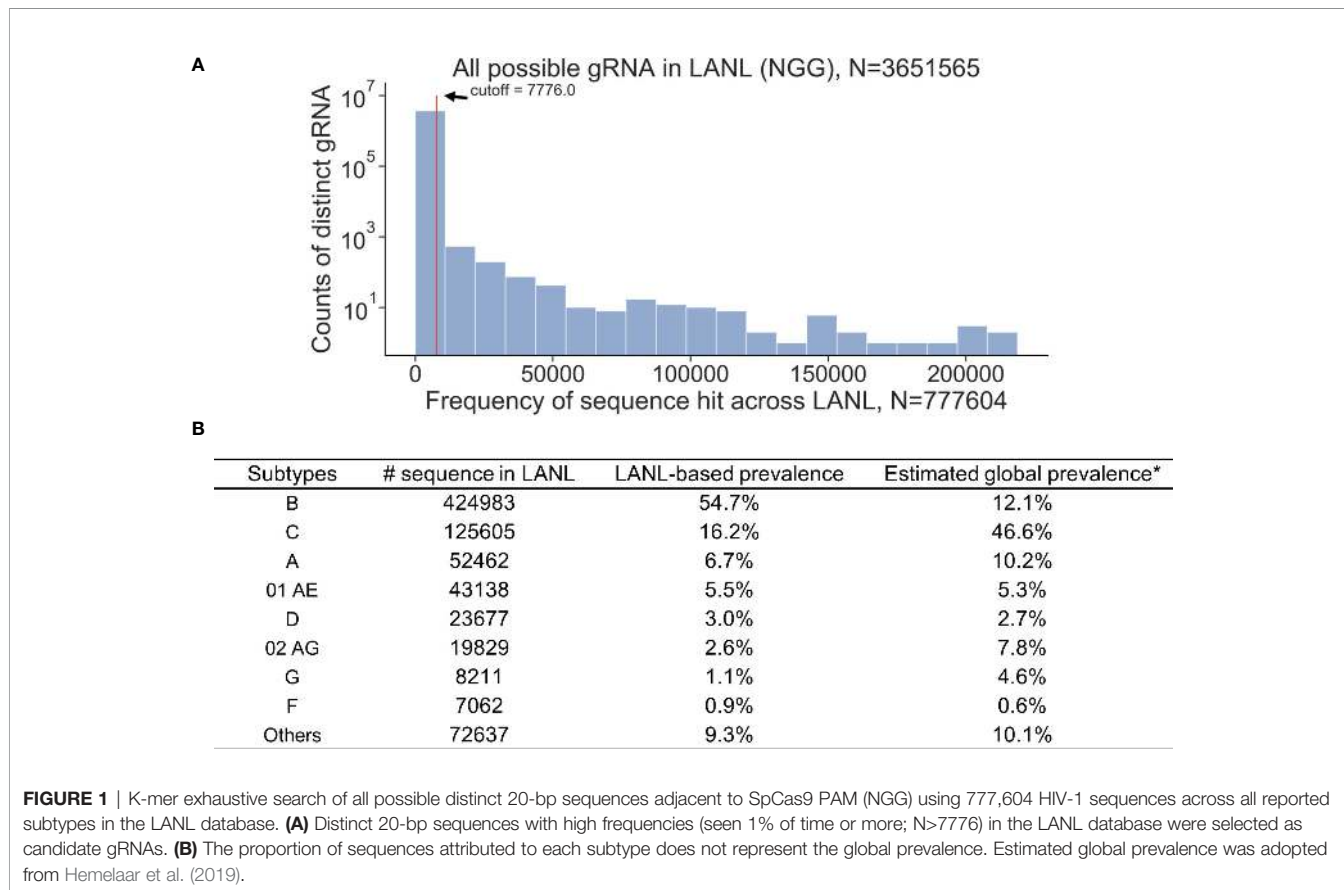
RESULTS

K-Mer-Based Search of Potential gRNAs in the LANL HIV Sequence Database

All available HIV-1 sequences deposited in the Los Alamos National Laboratory (LANL) database were used for the analysis (N=777,604, as of August 2018). A 3-bp PAM (NGG, any nucleotide followed by two consecutive guanines) required by Cas9 was the first restriction for gRNA design. An exhaustive search of all 20-mer adjacent to a SpCas9 PAM site were used to collect all possible gRNAs within the LANL database. This resulted in 3,651,565 distinct 20-bp sequences from either sense or antisense strands (**Figure 1A**). However, only 1.9% of possible gRNAs perfectly matched more than 100 sequences within the LANL database. gRNAs that perfectly matched at least 1% (7,776/777,604) of LANL sequences were considered high frequency gRNAs and used for subsequent analysis. Only 1,330 out of 3,651,565 candidate gRNAs exceeded this criterium (**Table S1**). The estimated global editing efficiency of candidate gRNAs was scaled at the population level by the estimated global prevalence for subsequent analysis to account for biased sample size toward subtype B in LANL database (**Figure 1B**). For example, subtype B represents 54.7% of the sequences but only 12.1% of the global prevalence (Hemelaar et al., 2019).
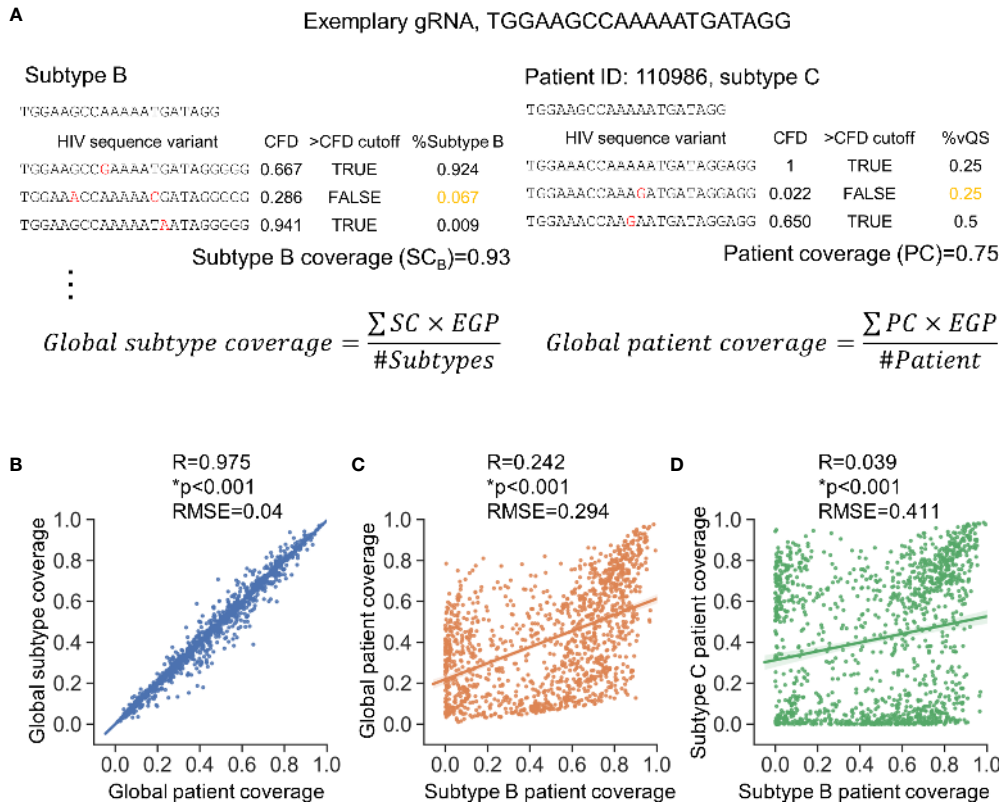
Cross-Subtype Estimation of gRNA Editing Efficiency

Two metrics (shown in **Figure 2A**) were used to assess anti-HIV-1 gRNAs across subtypes; global patient coverage and global subtype coverage. Global patient coverage utilizes data from the *Pat_SSID* column in LANL to group 777,604 sequences by individual "patients." The coverage for each patient was calculated and then averaged across the database (**Patient coverage section in Table S1**). Global Subtype Coverage was calculated by considering each of 777,604 sequences as an independent sample (**Subtype coverage section in Table S1**). The global coverage was then calculated by the summation of coverage by each subtype or each patient weighted by the subtype estimated global prevalence (EGP) (**Figure 2A**) (Robertson et al., 2000). The minor variants in global patient coverage is weighted more than that in global subtype coverage, which makes the global patient coverage a suitable metric that accounts for patients possessing only minor variants in latent reservoir for subsequent analyses.

There was a strong positive correlation between global subtype coverage and global patient coverage among the 1,330 candidate gRNAs (**Figure 2B**). Additional correlation tests were conducted to test how subtype coverage of gRNAs designed against Subtype B sequences specified in LANL HIV-1 sequence

Subtypes	# sequence in LANL	LANL-based prevalence	Estimated global prevalence*
B	424983	54.7%	12.1%
C	125605	16.2%	46.6%
A	52462	6.7%	10.2%
01 AE	43138	5.5%	5.3%
D	23677	3.0%	2.7%
02 AG	19829	2.6%	7.8%
G	8211	1.1%	4.6%
F	7062	0.9%	0.6%
Others	72637	9.3%	10.1%

FIGURE 1 | K-mer exhaustive search of all possible distinct 20-bp sequences adjacent to SpCas9 PAM (NGG) using 777,604 HIV-1 sequences across all reported subtypes in the LANL database. **(A)** Distinct 20-bp sequences with high frequencies (seen 1% of time or more; N>7776) in the LANL database were selected as candidate gRNAs. **(B)** The proportion of sequences attributed to each subtype does not represent the global prevalence. Estimated global prevalence was adopted from Hemelaar et al. (2019).

FIGURE 2 | Consideration of global genetic variants is necessary to identify high quality gRNAs. **(A)** Subtype coverage and patient coverage were defined to account for only subtype differences (left panel) or patient-specific differences (right panel), respectively. Subtype coverage was the summation of the frequency of variants that were predicted to be cut (>CFD cutoff) within subtype. Subtype coverage is the unit to account for global subtype coverage. On the other hand, patient coverage represented the summation of frequency of variants identified within patients that were predicted to be cut. The CFD cutoff here is 0.569 where the variants were predicted to be cut when CFD between testing gRNA and variant was larger than 0.569. EGP, Estimated global prevalence; SC, Subtype coverage; PC, Patient coverage. **(B)** Global subtype coverage and global patient coverage was weighted by estimated global prevalence. Both axes represent the percentage of patient infected by specific subtypes that could be treated by a given gRNA. Scale=[0,1]. **(C, D)** The subtype coverage calculated against subtype B is poorly correlated with the global coverage rate **(C)** and the subtype C coverage **(D)**. Correlation coefficient (R) was determined by Pearson correlation; *p<0.001, Wald test was used to test whether variable on the x-axis is a significant predictor of variable on the y-axis; RMSE, root-mean-square error.

database, where most research has been conducted, performed against global coverage. This analysis showed only moderate correlation (R^2 = 0.242) with a root-mean-squared error of 0.294 (**Figure 2C**). This modest correlation could also be observed between global patient coverage and the subtype-specific patient coverage in all other subtypes (**Figure S1**). Furthermore, inter-subtype patient coverages were even more diverse (**Figure 2D**, **Figure S1**). For example, the root-mean-squared error of patient coverages against subtype B and C among the 1,330 candidate gRNAs was 0.411 (R^2 = 0.039), indicating that gRNA design based on one subtype may have limited therapeutic potential for other subtypes.

Low Diversity Target Sites With High Global Patient Coverage gRNA Sequences Were Identified

Previous studies have shown that HIV-1 target sites with low genetic diversity significantly enhanced CRISPR-mediated HIV-1 inactivation efficiency and prevented CRISPR-induced viral escape (Wang et al., 2016a; Wang et al., 2016b; Lebbink et al., 2017; Mefferd et al., 2018). The genomic diversity of each subtype was estimated by calculating the cumulative 20-mer Shannon entropy after the sequences were aligned against HXB2 (Accession number K03455) as described in the *Methods*. The global sequence diversity was the estimated sum of the subtype-specific diversity weighted by the estimated global prevalence shown in **Figure 1B** (**Figure 3A**). The global diversity of target sites among the 1,330 candidate gRNAs ranged between 0.76 and 14.07 bits. Previous studies that have targeted low diversity regions within tat and rev exon 1 (gTatRev, target site:5970-5989) and gag p24 (gGag1, target site: 1389–1408) showed complete viral suppression with no sign of resistance over 110 days of culture (Wang et al., 2016a; Wang et al., 2016b; Darcis et al., 2019). Both were identified in 1,330 candidate gRNAs (g8D9BC2=gTatRev and g80892B=gGag1 in **Table S1**). The
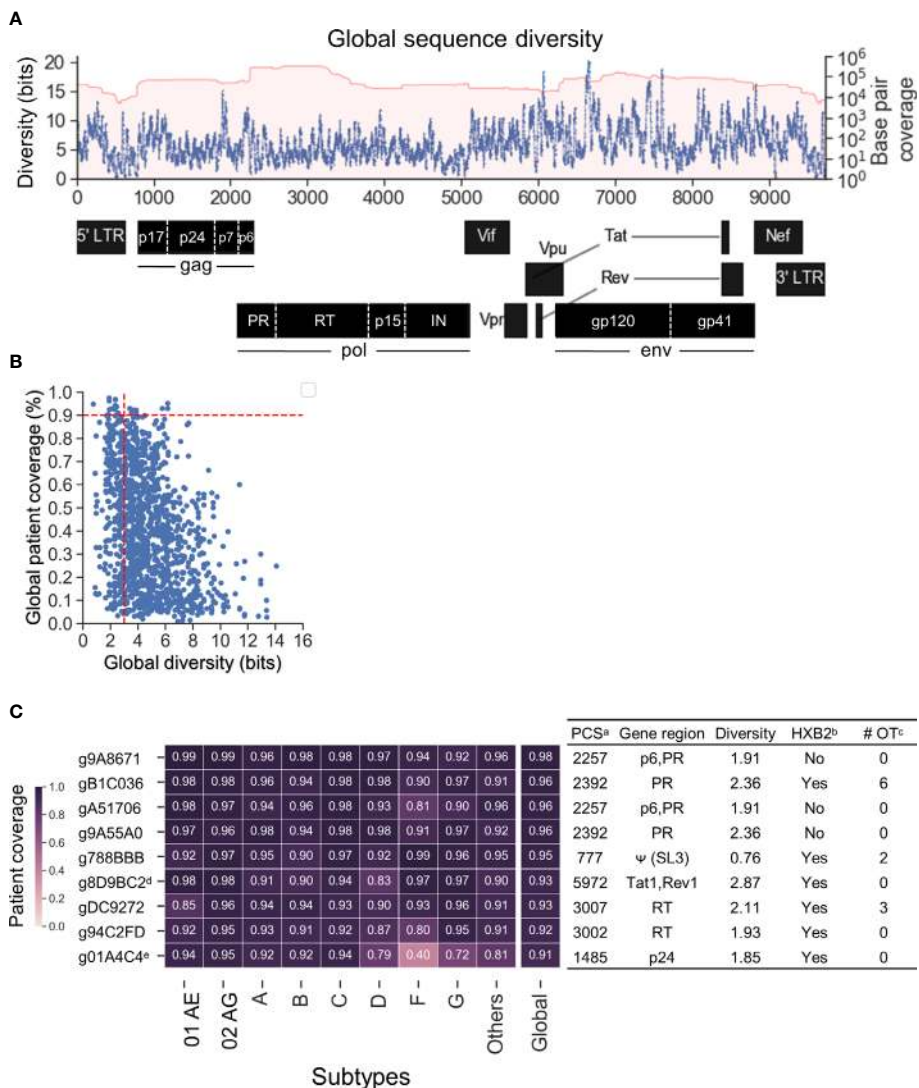
FIGURE 3 | Nine lead gRNAs were identified with high coverage across subtypes. **(A)** The map of global sequence diversity estimated by subtype diversity weighted by estimated global prevalence. **(B)** Scatterplot between global diversity and global patient coverage. Vertical dashed line showed a cutoff of 3 for global diversity and 0.9 for global patient coverage. Nine lead gRNAs were advanced by the cutoff selection. **(C)** The heatmap of subtype-specific patient coverage and global patient coverage. Of the nine gRNAs shown, two were previously published. The additional seven are novel gRNA sequences targeting five distinct HIV-1 genes/motifs as aligned against HXB2. [a]PCS: Predicted cleavage sites; [b]HXB2: Whether the absolute gRNA sequence perfectly matched with HXB2 reference sequence. Note that the NL4-3 and R7/E-/eGFP in J-Lat (Chung C.-H. et al., 2020b) also showed the same results as HXB2 (**Table S1**); c[#] OT: number of potential off-target sites with CFD above 0.569 in human genome. [d]g8D9BC2: Previously published as gTatRev in (Wang et al., 2016a; Wang et al., 2016b; Darcis et al., 2019); [e]g01A4C4: Previously published as gGag3 in (Wang et al., 2016a; Zhao et al., 2017).

global diversity in g8D9BC2 and g80892B was 2.87 and 3.01 respectively (**Table S1**). Candidate gRNAs were also scanned for off-target likelihood (**Table S1**) with representative profiles shown in **Figure S2**.

Given this observation, further gRNA selection criteria were set to a global diversity of 3.01 and global patient coverage to more than 90% (**Figure 3B**). Nine lead gRNAs passed the criteria. There were seven novel gRNAs identified which mainly targeted regions within gag and pol (**Figure 3C**). Two

of the nine gRNA sequences, g8D9BC2 (gTatRev) and g01A4C4 (gGag3), have been previously experimentally examined (Wang et al., 2016a; Wang et al., 2016b; Zhao et al., 2017; Darcis et al., 2019). Viral replication was reduced to 2.8% in four independent tests using g8D9BC2 measured by p24 level in the supernatant of culture compared to control. The use of g01A4C4 reduced p24 levels to 5.4% on average in two independent experiments. These observations indicated that the gRNA search method could reproduce previous

findings. Note again that this gRNA search method considers both sequence conservation and gRNA-target binding specificity. This means that a highly conserved region is not guaranteed to produce gRNAs with high patient or subtype coverage. For example, although target site 1389–1408 had a low global diversity at 3.01, g80892B (gGag1 in previous studies) was predicted to only cleave 55.5% of patient-derived HIV-1 variants (**Table S1**). More specific, g80892B covered only 52.1% of subtype C variants and 3% of subtype G. Six of nine gRNAs possessed identical sequences to subtype B molecular clones that were commonly used in the literature including HXB2, NL4-3, and R7/E-/eGFP in J-Lat (**Figure 3C**, **Table S1**) (Adachi et al., 1986; St Clair et al., 1991; Jordan et al., 2003). HXB2 was used as the representative for sequence variants that are identical to these reference sequences in subsequent analysis. The number of potential off-target sites was calculated based on the CFD matrix described in the *Methods* section and recorded in **Table S1**. The list of off-target sites and chromosomal positions was illustrated in **Figure S2**.

Overlapping High-Quality Candidate gRNAs Unraveled the Effect of Absolute Sequence Selection on Consequent CRISPR-Mediated Editing Efficacy

Overlapping candidate gRNAs were found for individual target gene regions (**Figure 3C**). All candidate gRNAs that overlapped with the nine lead gRNAs on the same target sites were then identified. The target sites g9A8671, gA51706, and g07CF4A overlapped at position 2255–2274 on the HIV-1 genome (**Figure 4A**). They possessed only one bp difference across the 20-bp target region among each other (**Figure 4A**). However, g07CF4A only had 16% global patient coverage (**Table S1**). Another proximal target site at position 2376–2395 had two overlapping gRNAs, g9A55A0 and gB1C036, that differed at position 6 of the 20-bp target site. (**Figure 4B**). Both target sites exhibited high conservation across subtypes (**Figure 4C**). The target site at position 2255–2274 covered both p6 and protease coding sequences, while the other targeted only the protease sequence. To explain these differences, the major sequence variants that represented up to 95% of all known variants were analyzed. g07CF4A held the highest CFD score (0.86) compared to the HXB2 sequence versus g9A8671 (0.61) and gA51706 (0.71) (**Figure 4D**). These scores were all higher than the CFD cutoff of 0.569 used to distinguish whether the gRNA was predicted to induce DSBs 95% of the time (Chung C-H et al., 2020). However, only 4.9% of variants in the LANL database possess an identical target site to HXB2 (**Figure 4D**). This emphasized the advantage of the design method presented here to target the most prevalent genetic variants in the known HIV-1 sequences. In other words, using one reference genome such as HXB2 or NL4-3 is not sufficient to find these gRNAs with high global patient coverage. Using this more generalized design method, gRNAs with high patient coverage could be identified.

Another interesting pair of gRNAs, gB1C036 and gRNA14, were both identified to target position 2376-2395 within HIV-

1 protease at high patient coverages (**Figure 4B**). While gB1C036 possessed an identical sequence against the most prevalent variant at this position (62.7%), gRNA14 only perfectly matched to 4.1% of variants in this region (**Figure 4E**). If a gRNA is not predicted to induce DSB on the predominant sequence variants (CFD score below 0.569) it loses the majority of patient coverage. However, the global patient coverage of g9A55A0 (96.4%) remained as high as that of gB1C036 (96.5%) (**Figure 4A**). This could be explained by the nucleotide difference at position six in CFD matrix. The guanine (G) on position six of g9A55A0 did not cause any penalty for the mismatch against adenine (A) on the target DNA. Furthermore, the overall CFD score was increased between g9A55A0 sequence and all major variants (**Figure 4E**).

Sequences Responsible for the Packaging Signal Are Under High Evolutionary Constraint

g788BBB was found to target a sequence region that should inactivate HIV-1 replication but has been previously uninvestigated (**Figure 5A**). This gRNA targets positions 761–780 across a sequence motif that forms one of the four functional RNA stem-loop secondary structures, conventionally termed SL3 (stem-loop 3) or Ψ, for nucleocapsid (p7) to bind during HIV-1 genome packaging (De Guzman et al., 1998). A previous study has showed lower sequence diversity within the SL3 region than adjacent genomic regions, indicating that mutations in this region are rarely tolerated (Ingemarsdotter et al., 2018). The predicted cleavage site was at position 777, which would likely induce InDels that disrupt the stem loop formation (**Figures 5B, C**). Furthermore, the PAM site was located 7-bp upstream of the translation initiation site (**Figure 5C**). More than 83.8% of sequences across subtypes were identical within the 20-bp target region, which was in agreement with previous observations. (**Figure 5D**) (Ingemarsdotter et al., 2018). This is the highest percentage of predominant variants covered among the nine lead gRNAs. Most of the minor variants existed in subtype 01 AE, B, and D indicated by the subtype-specific patient coverage (**Figure 3C**).

The Reverse Transcriptase-Targeting gRNA Covered a Conserved Subdomain Responsible for dNTP Incorporation

Target sites at positions 2991-3010 and 3000-3019 cover RT at residues 148-155 and 150-157, respectively (**Figures 6A, B**). gDC9272- and g94C2FD-targeted sites reside in the second finger subdomain in RT (**Figure 6C**). Previous studies have shown that the Q151 residue was responsible for direct interaction with the 3'-OH of the incoming dNTP (Huang et al., 1998). The Q151M mutation confers resistance against most nucleoside RT inhibitors (NRTIs), nucleoside analogs that lack the 3'-OH group (Ueno et al., 1995). The predicted cleavage site of g94C2FD was between residues 151–152, while gDC9272 was predicted to cleave between the second and third codon of
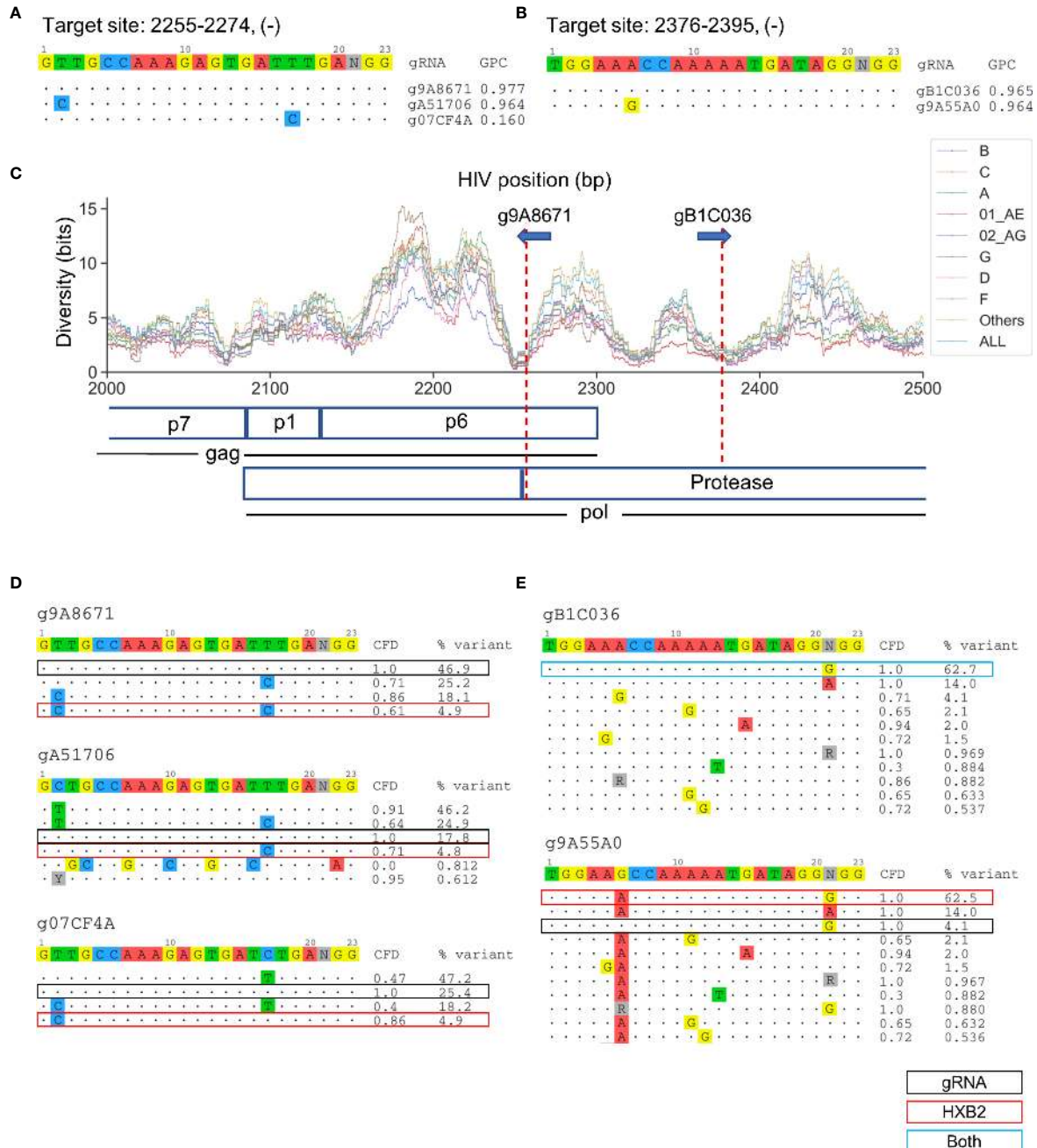
FIGURE 4 | The selection of absolute sequence is likely to affect consequent CRISPR-mediated editing efficacy across sequence variants. **(A)** All candidate gRNAs targeting position 2255–2274 (-) among 1,330 candidate gRNAs. **(B)** All candidate gRNAs targeting position 2376–2395 (-) among 1,330 candidate gRNAs. **(C)** The subtype-specific sequence diversity across the HIV-1 genome between position 2000 and 2500. Both g9A8671 and gB1C036 were predicted to cleave the protease protein in a locally low diversity region. gRNA589 was also predicted to cleave p6. **(D, E)** Sequence variant profiles and corresponding CFD score between chosen gRNA and HIV-1 variants. "HXB2" label next to the % variant represents the variant frequency of the sequence that possessed identical sequence against HXB2.

residue 153 (**Figure 6C**). gDC9272 was predicted to target 26/32 observed sequence variants listed in **Figure 6D**, which allowed for a 93.1% global variant coverage. A C-to-T mismatch at position 18 between the g94C2FD and the sequence variant

with 9.2% frequency resulted in moderate reduction of CFD score to 0.64 (**Figure 6E**). However, this reduction was not predicted to prevent CRISPR-mediated editing because the CFD score was still higher than the cutoff.
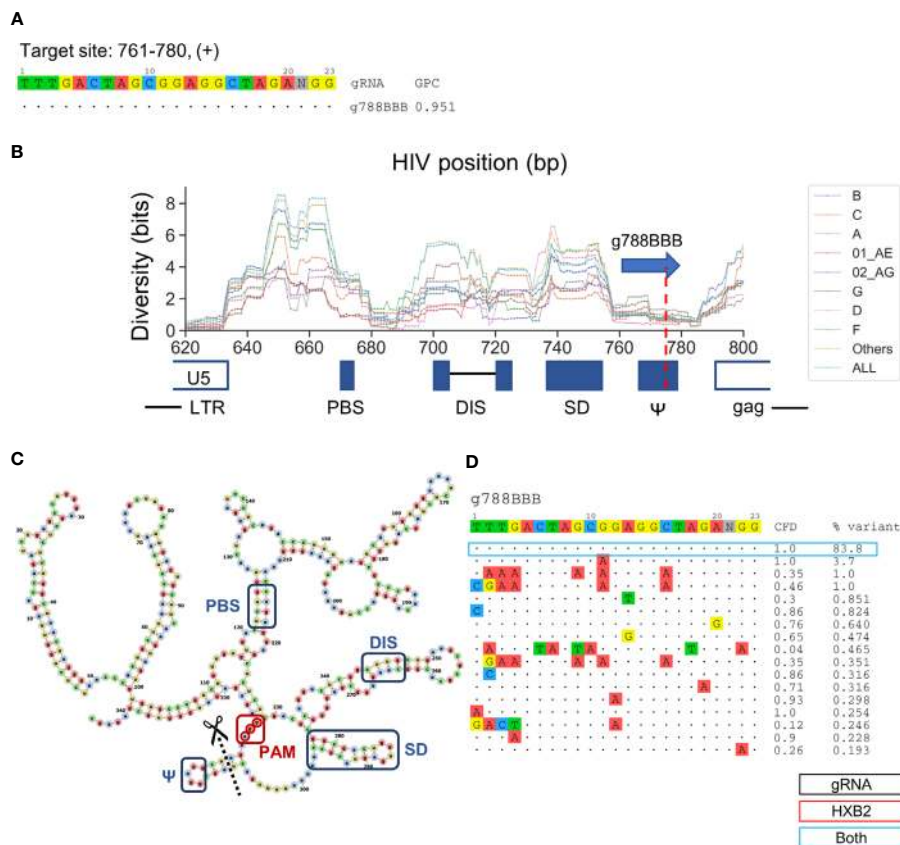
FIGURE 5 | Low diversity in the SL3 region is an ideal target for the selection of gRNA sequence that provides high predicted global patient coverage. **(A)** All candidate gRNAs targeting position 761-780 (+) among 1,330 candidate gRNAs. **(B)** The subtype-specific sequence diversity across the HIV-1 genome between position 600 and 800. The predicated target region showed a local trough flanking by g788BBB target sites. g788BBB was predicted to cleave nucleocapsid binding site, Ψ. **(C)** Predicted HXB2 untranslated region (UTR) secondary structure with the PBS (primer binding site), DIS (dimer initiation site), SD (splicing donor), and Ψ regions labeled with blue, g788BBB PAM labeled with red, and cleavage site labeled with black. Secondary structure was predicted using RNAfold (Gruber et al., 2008) and visualized using VARNA (Darty et al., 2009). **(D)** Sequence variant profiles and corresponding CFD score between chosen gRNA and HIV-1 variants. "HXB2" label next to the % variant represent the variant frequency of the sequence that possessed identical sequence against HXB2.

DISCUSSION

The overarching goal of this study was to generate broad-spectrum gRNAs for the generalized use of CRISPR-mediated antiviral therapy across patients infected by different subtypes. Multiple sequence alignments can produce spurious noise when aligning sequences with high diversity and numerous recombination events while a k-mer approach is unphased by this noise. In this study, the standard k-mer counting technique was supplemented with CRISPR-mediated editing mechanisms to optimize gRNA design methods. A pool of 20-bp candidate gRNAs that were conserved across sequence variants of all subtypes was selected. The data suggested that the consideration of subtype differences was essential during the design process. The gRNAs designed by incorporating sequence variants within single subtypes did not guarantee effectiveness across other subtypes (**Figure 2D** and **Figure S1**). A similar result was shown that previously proposed anti-HIV-1 gRNAs often lack homology to major sequence variants seen across infected patients (Dampier et al., 2017). The

metric used in this study to calculate global patient coverage and subtype coverage accounted for the difference of existing variation between subtypes.

Sequence variant coverage accounted for the likelihood a gRNA sequence that can cause DSBs on the collection of known variants. However, it did not estimate the functional relevance of targeted loci in the HIV-1 lifecycle. Previous studies have used sequence conservation to select HIV-1 target regions (Ebina et al., 2013; Liao et al., 2015; Ueda et al., 2016; Mefferd et al., 2018; Yin et al., 2018). The reduced sequence diversity was a consequence of low tolerance of mutations at functionally important domains. HIV-1 required a longer time to develop resistant strains when the gRNAs targeted more conserved regions (Wang et al., 2016b). A systematic approach that analyzed experimental readouts with the use of anti-HIV-1 gRNAs also demonstrated a positive correlation between functional reduction and sequence diversity at target sites. The low diversity regions identified in this study may point to sequence motifs with significant functions that were preserved
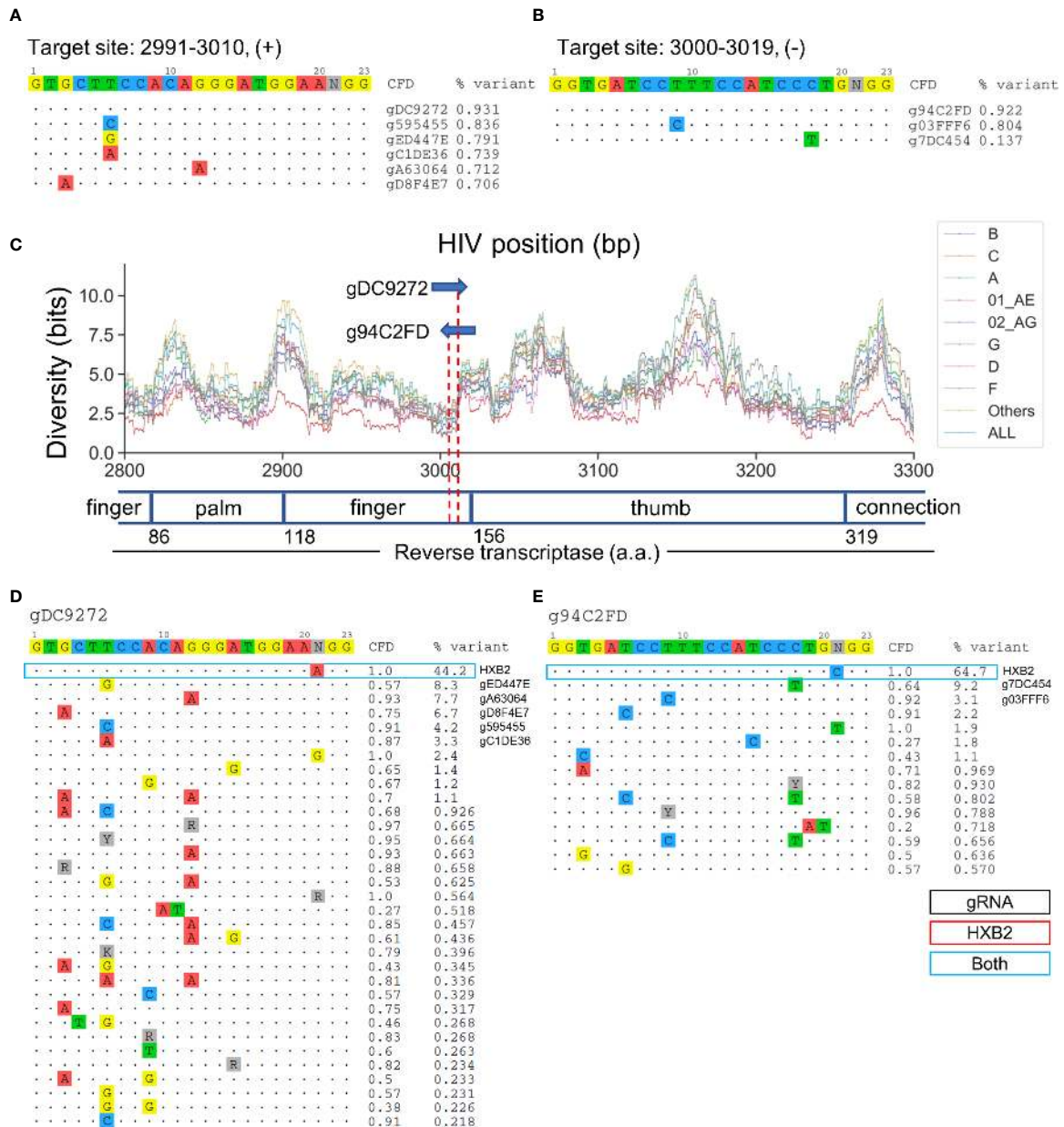
FIGURE 6 | Two lead gRNAs target critical dNTP binding domain in HIV RT protein that is responsible for effective reverse transcription. **(A)** All candidate gRNAs targeting position 2991-3010 (+) among 1,330 candidate gRNAs. **(B)** All candidate gRNAs targeting position 3000-3019 (+) among 1,330 candidate gRNAs. **(C)** The subtype-specific sequence diversity across the HIV-1 genome between position 2800 and 3300. Major domains are shown in the open or closed blocks. The numbers next to domain names indicate the amino acid coordinate for reverse transcriptase. **(D, E)** Sequence variant profiles and corresponding CFD score between chosen gRNA and HIV-1 variants. "HXB2" label next to the %variant represent the variant frequency of the sequence that possessed identical sequence compared with HXB2. The sequence variants labeled for gRNAs represent the gRNA contains the identical sequence with sequence variants.

in all HIV-1 subtypes. The final list of lead gRNAs was identified at the regions where the broad-spectrum sequence variants and low diversity sequence variants converged. This result made clear that the metrics optimizing absolute gRNA sequence and determining target sites agreed with each other better when the entropy was reduced. Note that three lead gRNAs identified in this study were predicted to have potential off-target cleavage

only at intergenic or intron regions. Further validation with respect to off-target cleavage using genome wide sequencing approach similar to previous study as well as *in vitro* cell viability test are required to support the safety of proposed gRNAs (Chung C.-H. et al., 2020a).

The gRNA search method identified nine gRNAs possessing global patient coverage of 0.9 or higher; seven of the novel gRNA

sequences that mapped to five distinct target sites have not been previously tested in the literature. Interestingly, the results presented in this study showed that the absolute sequence of gRNA and the target sequences across HIV-1 sequence variants largely affected the outcome of predicted patient coverage. All lead gRNAs were found to be identical to the most predominant variant at the target sites. An interesting result regarding g9A55A0 is that while it contained at least one mismatch to more than 96% of sequences, it still exhibited high patient coverage. It is possible that this phenomenon could be due to lack of experimental results that informed the relationship between this specific pairwise identity when the CFD matrix was derived. Functional studies are warranted to first distinguish whether the G-to-A mismatch at position 6 reduced editing efficiency against HXB2 or other molecular clones that contain the same target site sequence. In addition, this work demonstrated the computational strategy to identify optimal gRNAs targeting sites using the CFD matrix, which was derived from the data using SpCas9 system. This strategy could be adapted by other Cas orthologs. All potential gRNAs targeting HIV-1 LAI reference genome using SaCas9, nmCas9, and Cpf1 were listed in a previous study with little consideration given to HIV-1 genetic variation and subtype differences (Yoder, 2019). The same pipeline developed in this study could be utilized for gRNA screening and selection with different Cas systems if the position-specific penalty matrices are available. However, the CFD matrix was derived from functional study using only SpCas9. This means that a Cas ortholog might have an orthogonal penalty matrix that requires a systematic experimental design in independent studies. For example, a SpCas9 derivative such as SpCas9-HF1 or HiFi-SpCas9 may have similar behavior as SpCas9 but not identical (Kleinstiver et al., 2016; Vakulskas et al., 2018).

SL3 is one of the targetable sites that was previously unexamined in the test of HIV-1 inactivation efficiency using CRISPR/Cas. The 300-bp untranslated region (UTR) of the HIV-1 genome forms secondary RNA structures and plays a crucial role in the HIV-1 replication cycle. The transactivation response (TAR) element at the beginning of transcription has been found to be conserved and effective at reducing HIV-1 replication using TAR-targeting gRNAs (Hu et al., 2014; Kaminski et al., 2016b; Yin et al., 2016; Bella et al., 2018). However, other functional motifs in the UTR were rarely tested using the CRISPR system. A previous study examined gRNA gPBS1-3, which targeted a low diversity region located in primer binding site (PBS). The time of emerged escape mutants was delayed by gPBSs but did not circumvent later onset of resistance (Wang Z. et al., 2018). However, small molecules targeting secondary RNA structure in HIV-1 remain early in the developmental process for clinical use (Warui and Baranger, 2012; Ingemarsdotter et al., 2018). CRISPR-mediated editing is ideal for targeting non-coding functional domains since its mechanism of action is on the DNA level. This indicated a new modality to intervene in non-coding sequence regions that conventional biologics have found difficult to target. The characteristic of Ψ holds great promise for the use of CRISPR-mediated inactivation strategy due to low sequence diversity and

reduced number of distinct variants across all subtype present in the LANL database. It indicates that this region has been under high evolutionary pressure to reduce background mutations and to serve as an ideal target for gRNA design.

Overall, the novel target sites with optimal gRNA sequences were identified with promising characteristics of functional importance. Functional assays are warranted to validate the efficiency at reducing HIV-1 replication. A viral swarm derived from HIV-1-infected individuals was also proposed to take closer evaluation with respect to the ability of targeting genetic variants that were naturally occurred within and between patients.

DATA AVAILABILITY STATEMENT

The HIV-1 sequence datasets for this study can be found in the search interface in the LANL website [https://www.hiv.lanl.gov/components/sequence/HIV/search/search.html]. The analyses and figures can be reproduced in Python notebook (https://github.com/DamLabResources/HIV-subtype-gRNAs).

AUTHOR CONTRIBUTIONS

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2021.593077/full#supplementary-material

REFERENCES

Abecasis, A. B., Vandamme, A. M., and Lemey, P. (2009). Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J. Virol* 83, 12917–12924. doi: 10.1128/JVI.01022-09

Adachi, A., Gendelman, H. E., Koenig, S., Folks, T., Willey, R., Rabson, A., et al. (1986). Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *J. Virol* 59, 284–291. doi: 10.1128/JVI.59.2.284-291.1986

Barton, K., Winckelmann, A., and Palmer, S. (2016). HIV-1 Reservoirs During Suppressive Therapy. *Trends Microbiol.* 24, 345–355. doi: 10.1016/j.tim.2016.01.006

Bella, R., Kaminski, R., Mancuso, P., Young, W. B., Chen, C., Sariyer, R., et al. (2018). Removal of HIV DNA by CRISPR from Patient Blood Engrafts in Humanized Mice. *Mol. Ther. Nucleic Acids* 12, 275–282. doi: 10.1016/j.omtn.2018.05.021

Bui, J. K., Halvas, E. K., Fyne, E., Sobolewski, M. D., Koontz, D., Shao, W., et al. (2017). Ex vivo activation of CD4+ T-cells from donors on suppressive ART can lead to sustained production of infectious HIV-1 from a subset of infected cells. *PLoS Pathog.* 13, e1006230. doi: 10.1371/journal.ppat.1006230

Chung, C. H., Allen, A. G., Atkins, A. J., Sullivan, N. T., Homan, G., Costello, R., et al. (2020a). Safe CRISPR-Cas9 Inhibition of HIV-1 with High Specificity and Broad-Spectrum Activity by Targeting LTR NF-kappaB Binding Sites. *Mol. Ther. Nucleic Acids* 21, 965–982. doi: 10.1016/j.omtn.2020.07.016

Chung, C. H., Mele, A. R., Allen, A. G., Costello, R., Dampier, W., Nonnemacher, M. R., et al. (2020b). Integrated Human Immunodeficiency Virus Type 1 Sequence in J-Lat 10.6. *Microbiol. Resour. Announc.* 9 (18), e00179-20. doi: 10.1128/MRA.00179-20

Chung, C.-H., Allen, A. G., Atkins, A. J., Sullivan, N. T., Homan, G., Costello, R., et al. (2020). Safe CRISPR/Cas9 inactivation of HIV-1 transcription with high specificity and broad-spectrum activity in latently infected cells by mutation of HIV-1 promoter NF-κB binding sites. *Mol. Therapy-Nucleic Acids* 21, 965–982. doi: 10.1016/j.omtn.2020.07.016

Colby, D. J., Trautmann, L., Pinyakorn, S., Leyre, L., Pagliuzza, A., Kroon, E., et al. (2018). Rapid HIV RNA rebound after antiretroviral treatment interruption in persons durably suppressed in Fiebig I acute HIV infection. *Nat. Med.* 24, 923–926. doi: 10.1038/s41591-018-0026-6

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823. doi: 10.1126/science.1231143

Cui, Y., Xu, J., Cheng, M., Liao, X., and Peng, S. (2018). Review of CRISPR/Cas9 sgRNA Design Tools. *Interdiscip. Sci.* 10, 455–465. doi: 10.1007/s12539-018-0298-z

Dampier, W., Nonnemacher, M. R., Sullivan, N. T., Jacobson, J. M., and Wigdahl, B. (2014). HIV Excision Utilizing CRISPR/Cas9 Technology: Attacking the Proviral Quasispecies in Reservoirs to Achieve a Cure. *MOJ Immunol.* 1 (4), 00022. doi: 10.15406/moji.2014.01.00022

Dampier, W., Nonnemacher, M. R., Mell, J., Earl, J., Ehrlich, G. D., Pirrone, V., et al. (2016). HIV-1 Genetic Variation Resulting in the Development of New Quasispecies Continues to Be Encountered in the Peripheral Blood of Well-Suppressed Patients. *PLoS One* 11, e0155382. doi: 10.1371/journal.pone.0155382

Dampier, W., Sullivan, N. T., Chung, C. H., Mell, J. C., Nonnemacher, M. R., and Wigdahl, B. (2017). Designing broad-spectrum anti-HIV-1 gRNAs to target patient-derived variants. *Sci. Rep.* 7, 14413. doi: 10.1038/s41598-017-12612-z

Dampier, W., Sullivan, N. T., Mell, J. C., Pirrone, V., Ehrlich, G. D., Chung, C. H., et al. (2018a). Broad-Spectrum and Personalized Guide RNAs for CRISPR/Cas9 HIV-1 Therapeutics. *AIDS Res. Hum. Retroviruses* 34, 950–960. doi: 10.1089/aid.2017.0274

Dampier, W., Chung, C. H., Sullivan, N. T., Atkins, A., Nonnemacher, M. R., and Wigdahl, B. (2018b). CRSeek: a Python module for facilitating complicated CRISPR design strategies. *PeerJ.* 6, e27094v1. doi: 10.7287/peerj.preprints.27094v1

Darcis, G., Binda, C. S., Klaver, B., Herrera-Carrillo, E., Berkhout, B., and Das, A. T. (2019). The Impact of HIV-1 Genetic Diversity on CRISPR-Cas9 Antiviral Activity and Viral Escape. *Viruses* 11 (3), 255. doi: 10.3390/v11030255

Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25, 1974–1975. doi:10.1093/bioinformatics/btp250

Dash, P. K., Kaminski, R., Bella, R., Su, H., Mathews, S., Ahooyi, T. M., et al. (2019). Sequential LASER ART and CRISPR Treatments Eliminate HIV-1 in a Subset of Infected Humanized Mice. *Nat. Commun.* 10, 2753. doi: 10.1038/s41467-019-10366-y

De Guzman, R. N., Wu, Z. R., Stalling, C. C., Pappalardo, L., Borer, P. N., and Summers, M. F. (1998). Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. *Science* 279, 384–388. doi: 10.1126/science.279.5349.384

De Leys, R., Vanderborght, B., Vanden Haesevelde, M., Heyndrickx, L., van Geel, A., Wauters, C., et al. (1990). Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin. *J. Virol.* 64, 1207–1216. doi: 10.1128/JVI.64.3.1207-1216.1990

de Silva, T. I., Cotten, M., and Rowland-Jones, S. L. (2008). HIV-2: the forgotten AIDS virus. *Trends Microbiol.* 16, 588–595. doi: 10.1016/j.tim.2008.09.003

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* 34, 184–191. doi: 10.1038/nbt.3437

Ebina, H., Misawa, N., Kanemura, Y., and Koyanagi, Y. (2013). Harnessing the CRISPR/Cas9 system to disrupt latent HIV-1 provirus. *Sci. Rep.* 3, 2510. doi: 10.1038/srep02510

Fauci, A. S. (2017). An HIV Vaccine Is Essential for Ending the HIV/AIDS Pandemic. *JAMA* 318, 1535–1536. doi: 10.1001/jama.2017.13505

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R., and Hofacker, I. L. (2008). The Vienna RNA websuite. *Nucleic Acids Res.* 36, W70–W74. doi: 10.1093/nar/gkn188

Haeussler, M., Schonig, K., Eckert, H., Eschstruth, A., Mianne, J., Renaud, J. B., et al. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* 17, 148. doi: 10.1186/s13059-016-1012-2

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2019). PD Ghys, and W-UNfHI Characterisation. Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. *Lancet Infect. Dis.* 19, 143–155. doi: 10.1016/S1473-3099(18)30647-9

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* 18, 182–192. doi: 10.1016/j.molmed.2011.12.001

Henrich, T. J., Hatano, H., Bacon, O., Hogan, L. E., Rutishauser, R., Hill, A., et al. (2017). HIV-1 persistence following extremely early initiation of antiretroviral therapy (ART) during acute HIV-1 infection: An observational study. *PLoS Med.* 14, e1002417. doi: 10.1371/journal.pmed.1002417

Hosmane, N. N., Kwon, K. J., Bruner, K. M., Capoferri, A. A., Beg, S., Rosenbloom, D. I., et al. (2017). Proliferation of latently infected CD4(+) T cells carrying replication-competent HIV-1: Potential role in latent reservoir dynamics. *J. Exp. Med.* 214, 959–972. doi: 10.1084/jem.20170193

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* 31, 827–832. doi: 10.1038/nbt.2647

Hu, W. S., and Hughes, S. H. (2012). HIV-1 reverse transcription. *Cold Spring Harb Perspect. Med.* 2 (10), a006882. doi: 10.1101/cshperspect.a006882

Hu, W., Kaminski, R., Yang, F., Zhang, Y., Cosentino, L., Li, F., et al. (2014). RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. *Proc. Natl. Acad. Sci. U. S. A.* 111, 11461–11466. doi: 10.1073/pnas.1405186111

Huang, H., Chopra, R., Verdine, G. L., and Harrison, S. C. (1998). Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* 282, 1669–1675. doi: 10.1126/science.282.5394.1669

Ingemarsdotter, C. K., Zeng, J., Long, Z., Lever, A. M. L., and Kenyon, J. C. (2018). An RNA-binding compound that stabilizes the HIV-1 gRNA packaging signal structure and specifically blocks HIV-1 RNA encapsidation. *Retrovirology* 15, 25. doi: 10.1186/s12977-018-0407-4

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. *Elife* 2, e00471. doi: 10.7554/eLife.00471

Jordan, A., Bisgrove, D., and Verdin, E. (2003). HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J.* 22, 1868–1877. doi: 10.1093/emboj/cdg188

Kaminski, R., Bella, R., Yin, C., Otte, J., Ferrante, P., Gendelman, H. E., et al. (2016a). Excision of HIV-1 DNA by gene editing: a proof-of-concept in vivo study. *Gene Ther.* 23, 690–695. doi: 10.1038/gt.2016.41

Kaminski, R., Chen, Y., Fischer, T., Tedaldi, E., Napoli, A., Zhang, Y., et al. (2016b). Elimination of HIV-1 Genomes from Human T-lymphoid Cells by CRISPR/Cas9 Gene Editing. *Sci. Rep.* 6, 22555. doi: 10.1038/srep22555

Kearney, M., Maldarelli, F., Shao, W., Margolick, J. B., Daar, E. S., Mellors, J. W., et al. (2009). Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *J. Virol.* 83, 2715–2727. doi: 10.1128/JVI.01960-08

Kearney, M. F., Spindler, J., Shao, W., Yu, S., Anderson, E. M., O'Shea, A., et al. (2014). Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS Pathog.* 10, e1004010. doi: 10.1371/journal.ppat.1004010

Keele, B. F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., et al. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 313, 523–526. doi: 10.1126/science.1126531

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z., et al. (2016). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* 529, 490–495. doi: 10.1038/nature16526

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours. Evolutionary, V. (2001). and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull.* 58, 19–42. doi: 10.1093/bmb/58.1.19

Lebbink, R. J., de Jong, D. C., Wolters, F., Kruse, E. M., van Ham, P. M., Wiertz, E. J., et al. (2017). A combinational CRISPR/Cas9 gene-editing approach can halt HIV replication and prevent viral escape. *Sci. Rep.* 7, 41968. doi: 10.1038/srep41968

Liao, H. K., Gu, Y., Diaz, A., Marlett, J., Takahashi, Y., Li, M., et al. (2015). Use of the CRISPR/Cas9 system as an intracellular defense against HIV-1 infection in human cells. *Nat. Commun.* 6, 6413. doi: 10.1038/ncomms7413

Maldarelli, F., Palmer, S., King, M. S., Wiegand, A., Polis, M. A., Mican, J., et al. (2007). ART suppresses plasma HIV-1 RNA to a stable set point predicted by pretherapy viremia. *PLoS Pathog.* 3, e46. doi: 10.1371/journal.ppat.0030046

Mefferd, A. L., Bogerd, H. P., Irwan, I. D., and Cullen, B. R. (2018). Insights into the mechanisms underlying the inactivation of HIV-1 proviruses by CRISPR/Cas. *Virology* 520, 116–126. doi: 10.1016/j.virol.2018.05.016

Ophinni, Y., Inoue, M., Kotaki, T., and Kameoka, M. (2018). CRISPR/Cas9 system targeting regulatory genes of HIV-1 inhibits viral replication in infected T-cell cultures. *Sci. Rep.* 8, 7784. doi: 10.1038/s41598-018-26190-1

Palmer, S., Maldarelli, F., Wiegand, A., Bernstein, B., Hanna, G. J., Brun, S. C., et al. (2008). Low-level viremia persists for at least 7 years in patients on suppressive antiretroviral therapy. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3879–3884. doi: 10.1073/pnas.0800050105

Persaud, D., and Luzuriaga, K. (2014). Absence of HIV-1 after treatment cessation in an infant. *N. Engl. J. Med.* 370, 678. doi: 10.1056/NEJMc1315498

Phillips, R. E., Rowland-Jones, S., Nixon, D. F., Gotch, F. M., Edwards, J. P., Ogunlesi, A. O., et al. (1991). Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* 354, 453–459. doi: 10.1038/354453a0

Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M., and Holmes, E. C. (2001). Human immunodeficiency virus. Phylogeny and the origin of HIV-1. *Nature* 410, 1047–1048. doi: 10.1038/35074179

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., et al. (2000). HIV-1 nomenclature proposal. *Science* 288, 55–56. doi: 10.1126/science.288.5463.55d

Roques, P., Robertson, D. L., Souquiere, S., Apetrei, C., Nerrienet, E., Barre-Sinoussi, F., et al. (2004). Phylogenetic characteristics of three new HIV-1 N strains and implications for the origin of group N. *AIDS* 18, 1371–1381. doi: 10.1097/01.aids.0000125990.86904.28

Shah, S. K., Persaud, D., Wendler, D. S., Taylor, H. A., Gay, H., Kruger, M., et al. (2014). Research on very early ART in neonates at risk of HIV infection. *Lancet Infect. Dis.* 14, 797. doi: 10.1016/S1473-3099(14)70893-X

Simon, F., Mauclere, P., Roques, P., Loussert-Ajaka, I., Muller-Trutwin, M. C., Saragosti, S., et al. (1998). Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat. Med.* 4, 1032–1037. doi: 10.1038/2017

St Clair, M. H., Martin, J. L., Tudor-Williams, G., Bach, M. C., Vavro, C. L., King, D. M., et al. (1991). Resistance to ddI and sensitivity to AZT induced by a mutation in HIV-1 reverse transcriptase. *Science* 253, 1557–1559. doi: 10.1126/science.1716788

Sturdevant, C. B., Joseph, S. B., Schnell, G., Price, R. W., Swanstrom, R., and Spudich, S. (2015). Compartmentalized replication of R5 T cell-tropic HIV-1 in the central nervous system early in the course of infection. *PLoS Pathog.* 11, e1004720. doi: 10.1371/journal.ppat.1004720

Sullivan, N. T., Dampier, W., Chung, C. H., Allen, A. G., Atkins, A., Pirrone, V., et al. (2019). Novel gRNA design pipeline to develop broad-spectrum CRISPR/Cas9 gRNAs for safe targeting of the HIV-1 quasispecies in patients. *Sci. Rep.* 9, 17088. doi: 10.1038/s41598-019-52353-9

Ueda, S., Ebina, H., Kanemura, Y., Misawa, N., and Koyanagi, Y. (2016). Anti-HIV-1 potency of the CRISPR/Cas9 system insufficient to fully inhibit viral replication. *Microbiol. Immunol.* 60, 483–496. doi: 10.1111/1348-0421.12395

Ueno, T., Shirasaka, T., and Mitsuya, H. (1995). Enzymatic characterization of human immunodeficiency virus type 1 reverse transcriptase resistant to multiple 2′,3′-dideoxynucleoside 5′-triphosphates. *J. Biol. Chem.* 270, 23605–23611. doi: 10.1074/jbc.270.40.23605

Vakulskas, C. A., Dever, D. P., Rettig, G. R., Turk, R., Jacobi, A. M., Collingwood, M. A., et al. (2018). A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nat. Med.* 24, 1216–1224. doi: 10.1038/s41591-018-0137-0

Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., et al. (2011). Confirmation of putative HIV-1 group P in Cameroon. *J. Virol.* 85, 1403–1407. doi: 10.1128/JVI.02005-10

Van Heuverswyn, F., Li, Y., Bailes, E., Neel, C., Lafay, B., Keele, B. F., et al. (2007). and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. *Virology* 368, 155–171. doi: 10.1016/j.virol.2007.06.018

Wang, G., Zhao, N., Berkhout, B., and Das, A. T. (2016a). A Combinatorial CRISPR-Cas9 Attack on HIV-1 DNA Extinguishes All Infectious Provirus in Infected T Cell Cultures. *Cell Rep.* 17, 2819–2826. doi: 10.1016/j.celrep.2016.11.057

Wang, G., Zhao, N., Berkhout, B., and Das, A. T. (2016b). CRISPR-Cas9 Can Inhibit HIV-1 Replication but NHEJ Repair Facilitates Virus Escape. *Mol. Ther.* 24, 522–526. doi: 10.1038/mt.2016.24

Wang, Q., Liu, S., Liu, Z., Ke, Z., Li, C., Yu, X., et al. (2018). Genome scale screening identification of SaCas9/gRNAs for targeting HIV-1 provirus and suppression of HIV-1 infection. *Virus Res.* 250, 21–30. doi: 10.1016/j.virusres.2018.04.002

Wang, Z., Wang, W., Cui, Y. C., Pan, Q., Zhu, W., Gendron, P., et al. (2018). HIV-1 Employs Multiple Mechanisms To Resist Cas9/Single Guide RNA Targeting the Viral Primer Binding Site. *J. Virol.* 92 (20), e01135-18. doi: 10.1128/JVI.01135-18

Warui, D. M., and Baranger, A. M. (2012). Identification of small molecule inhibitors of the HIV-1 nucleocapsid-stem-loop 3 RNA complex. *J. Med. Chem.* 55, 4132–4141. doi: 10.1021/jm2007694

Wei, X., Decker, J. M., Wang, S., Hui, H., Kappes, J. C., Wu, X., et al. (2003). Antibody neutralization and escape by HIV-1. *Nature* 422, 307–312. doi: 10.1038/nature01470

World Health Organization. (2020) *Global Health Observatory data repository.* Available at: https://www.who.int/docs/default-source/hiv-hq/latest-hiv-estimates-and-updates-on-hiv-policies-uptake-november2020.pdf?sfvrsn=10a0043d_12.

Yin, C., Zhang, T., Li, F., Yang, F., Putatunda, R., Young, W. B., et al. (2016). Functional screening of guide RNAs targeting the regulatory and structural HIV-1 viral genome for a cure of AIDS. *AIDS* 30, 1163–1174. doi: 10.1097/QAD.0000000000001079

Yin, C., Zhang, T., Qu, X., Zhang, Y., Putatunda, R., Xiao, X., et al. (2017). In Vivo Excision of HIV-1 Provirus by saCas9 and Multiplex Single-Guide RNAs in Animal Models. *Mol. Ther.* 25, 1168–1186. doi: 10.1016/j.ymthe.2017.03.012

Yin, L., Hu, S., Mei, S., Sun, H., Xu, F., Li, J., et al. (2018). CRISPR/Cas9 Inhibits Multiple Steps of HIV-1 Infection. *Hum. Gene Ther.* 29, 1264–1276. doi: 10.1089/hum.2018.018

Yoder, K. E. (2019). A CRISPR/Cas9 library to map the HIV-1 provirus genetic fitness. *Acta Virol* 63, 129–138. doi: 10.4149/av_2019_201

Yu, Q., Konig, R., Pillai, S., Chiles, K., Kearney, M., Palmer, S., et al. (2004). Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nat. Struct. Mol. Biol.* 11, 435–442. doi: 10.1038/nsmb758

Zanini, F., Brodin, J., Thebo, L., Lanz, C., Bratt, G., Albert, J., et al. (2015). Population genomics of intrapatient HIV-1 evolution. *Elife* 4, e11282. doi: 10.7554/eLife.11282

Zhao, N., Wang, G., Das, A. T., and Berkhout, B. (2017). Combinatorial CRISPR-Cas9 and RNA Interference Attack on HIV-1 DNA and RNA Can Lead to Cross-Resistance. *Antimicrob. Agents Chemother.* 61 (12), e01486-17. doi: 10.1128/AAC.01486-17

Zhu, W., Lei, R., Le Duff, Y., Li, J., Guo, F., Wainberg, M. A., et al. (2015). The CRISPR/Cas9 system inactivates latent HIV-1 proviral DNA. *Retrovirology* 12, 22. doi: 10.1186/s12977-015-0150-z