PRESENTATION PAGE

ORIGINAL ARTICLE

TITLE

shRNA design based on natural and drug induced variability targeting conserved regions from HIV reverse transcriptase

RUNNING TITLE

shRNAs directed against reverse transcriptase of HIV

COMPLETE AUTHORS INFORMATION

María Catalina Méndez Ortega[1, 2]

Juan Carlos Mendoza Sánchez[3]

Luis Miguel Rodríguez Rojas[2]

Silvia Restrepo[2]

Gloria Janeth Rey Benito[1]

[1]Grupo de Virología SRNL del Instituto Nacional de Salud, Bogotá, Colombia

[2]Laboratorio de Micología y Fitopatología de la Universidad de los Andes, Bogotá, Colombia

[3]VC@SOFT, Bogotá, Colombia

Correspondence should be addressed to M.C.M.

mc.mendez159@uniandes.edu.co.

Corresponding author:

María Catalina Méndez Ortega, Grupo de Virología, Instituto Nacional de Salud, Avenida Calle 26 No. 51-20, Bogotá, D.C., Laboratorio de Micología y Fitopatología de la Universidad de los Andes (LAMFU)

Telephone: (1) 2207700 ext 439 - 341; fax number: (1) 2207700 ext 340

e-mail: mc.mendez@uniandes.edu.co

ABSTRACT

Inhibition of HIV-1 by RNA interference can be achieved by the expression of short hairpin RNAs (shRNAs) targeting conserved sequences. We designed shRNAs against HIV-1 reverse transcriptase considering the variability of conserved regions within the retrovirus reverse transcriptase conserved domain cd01645, where the RNA dependent DNA polymerase activity of the enzyme resides. HXB2 (K03455) genome was used to choose regions with conserved residues important for enzyme activity. Regions were identified in multiple alignments from naive and drug-resistant isolates. A script was developed that predicted silencing efficiency based on previously reported parameters and that could identify highly frequent nucleotide combinations in a 21 base-long window within these regions. Higher number of sequence combinations was found in alignments from resistant isolates. shRNAs targeting reverse transcriptase active site residues W24 and P25 had scores over 7. Three-dimensional structural analyses from wild-type and resistant enzymes consistent with enormous variability explain the difficulties in finding a perfect region for RNAi mediated silencing. For clinical purposes, HIV-1 variability is an obstacle for efficient silencing from shRNAs designed towards a consensus sequence as there are too many functional variants of the enzyme. We suggest a complementary therapy using a combinatorial approach consisting of a mix of shRNAs targeting the most frequent viral variants from 1214 naïve genomes and 1381 from resistant variants, during a highly active antiretroviral therapy regimen.

Key Words: Acquired Immunodeficiency Syndrome, HIV-1, RNA interference, Gene Therapy, Antiretroviral Therapy, HAART

INTRODUCTION

Despite the advent of highly active antiretroviral therapy (HAART), human immunodeficiency virus (HIV-1) is still matter of public health concern [1, 2]. The major obstacle towards a cure lies in the integration of the viral genome [3-7]. Consequently, the source of variability exists indefinitely and the virus will always have the chance to restart infection and become resistant leading to treatment failure and AIDS [8-13].

The overwhelming genetic variability in HIV-1 is principally due to the error-prone nature of reverse transcriptase (RT), characterized by its low fidelity with a calculated incorporation of one mutation per replication cycle [15, 22]. But other three mechanisms are also important: i) homologous and non-homologous recombination events between the two copies of the viral genome incorporated within the virion [14-17], ii) reverse transcriptase (RT) translocations during reverse transcription [18-21], and iii) high replication rate of the virus, enough to produce 10^{10} viral particles per day. These factors are responsible for the generation of quasiespecies, that in conjunction with other forces contribute to the immune system exhaustion [6, 11, 23-26]. In addition, the virus source hides inside cellular reservoirs [8, 13, 27-29], which adds to its highly polymorphic nature, to its innate ability to accumulate mutations and to the flexibility of its proteins to accept them. Altogether these factors explain how the virus manages to overcome HAART [30, 31]. Clearly, therapy or vaccine strategies should be efficient against each and

every possible competent-replication variant that may emerge even in presence of selective pressures different from the immune system [32-34]. Unfortunately, though HAART saves thousands of lives, emergence of resistant variants occurs, even though multiple key steps of the replication cycle of the virus are targeted simultaneously [9, 35-37].

RNA interference (RNAi) is an eukaryotic evolutionarily conserved natural occurring process by which double-stranded RNA (dsRNA) triggers post-transcriptional gene silencing [38]. Its existence is related to a defense mechanism against infectious or foreign nucleic acids (like viruses), to the control of mobile genetic elements, and to the regulation of gene expression [39].

RNAi is nowadays considered an appropriate alternative to HAART nonetheless; efficacy may be lowered in humans due to viral variability and resistance emergence. Several RNAi *in vitro* approaches have proven to be effective against HIV-1 [32, 40-45]. Furthermore, HIV-1 was silenced *in vivo* in humanized mice models [40, 46-49]. Certainly, a potent replication inhibition was obtained by the simultaneous targeting of multiple conserved regions in the viral genome with expression of multiple shRNAs [45, 50-53]. HIV variability underlies the fact that key target selection is of utmost importance. As an example, it was previously shown *in vitro* that when a sequence located within the *nef* gene, which is present in all the viral mRNA species, is targeted, a region including the sequence is deleted from the viral genome rendering the virus resistant to RNAi silencing [116]. This happens because that specific sequence is biologically unimportant, and this cannot be ignored even though the target was not thought to be Nef itself. Up to now shRNA design has been based on conserved sequences

from a reference sequence, on a multiple alignment and in some cases from previously in vitro screened silencing molecules from which the most efficient ones have been selected [50, 52, 54]. ter Brake et al 2006, found 19 highly conserved regions that were shared by 75% of 170 genomes [45] and in a similar approach Naito et al, found 216 sequences that showed conservation in more than 75% of a total of 495 nearly full-length genomes [55]. The most recent approach analyzed 37,949 gene sequence fragments, from more than 8,000 genomes [56]. But no strategy has tried to understand and decipher HIV variability in terms of not trying to find a unique sequence for a specified conserved region, but instead in terms of identifying a region that has the least number of sequence combinations or viral variants from which the most frequent are chosen in order to design silencing molecules toward them. HIV variability makes impossible to have a unique sequence to target all the variants or quiasispecies that can exist.

So even though silencing of HIV in humanized mice models undoubtedly showed the potential of RNAi for HIV-1 treatment, it is still far from being effective as unique therapy in an infected patient [46, 47, 49, 57-59]. HIV-1 mutates within the constraints of the protein structure-function relationship and a mouse life span is perhaps too short to evidence the natural outcome of resistant quasispecies. Resistance in humans against HAART inhibitors may arise within a few or more years, usually 2 to 4, of initiating therapy depending on different variables [8, 9, 60-62]. Then, longer periods of time are needed in an animal model in order to evaluate emergence of resistance or the inhibition of virus replication, even more when treatment has to be life-long. Unquestionably, shRNAs must be carefully selected [63].

We propose a shRNA design approach considering all the variability from HIV sequences available in non-redundant HIV databases. We aimed to find a conserved region in which few shRNAs can be used to target all the sequences in a multiple alignment, using a Score-based attempt in order to aid selecting for efficient molecules. Our approach is based on identifying all the nucleotide changes that may be accepted in conserved 21-nucleotide long regions localized within conserved domains of HIV-1 RT. For this, we aimed to detect changes occurring during selective drug pressure by analyzing multiple sequence alignments performed with resistant isolates sequences retrieved from HIV Drug Resistance Database [64-66], based on a first line RT inhibitor (RTI) treatment. A curated multiple sequence alignment from HIV databases was also used to search for natural occurring changes. As nucleotide changes cannot accumulate in an infinite fashion and have to follow structure restrictions to preserve enzyme function we hypothesized that a set of shRNAs might exist that can silence specific virus variants and possibly all of them. Our main target is the RT conserved motif YMDD [67] that lies within the retrovirus reverse transcriptase conserved domain cd01645 RT_Rtv [67, 68] which is responsible for the RNA dependent DNA polymerase activity, required for the viral replication cycle [67, 69] and thus we expect the least number of changes to be allowed.

MATERIALS AND METHODS

Identification of conserved regions and possible target sites within the RT_Rtv domain

The search for conserved domains was performed in Pfam and InterProScan databases [70-74]. Identified domains were then localized in the reference sequence HXB2 accession number K03455 (gi|1906382|gb|K03455.1|HIVHXB2CG Human immunodeficiency virus type 1 (HXB2), complete genome; HIV1/HTLV-III/LAV reference genome). This sequence was used as query against the conserved domain database of NCBI [75], in order to identify catalytic residues, residues involved in DNA binding, dNTP binding, or active site residues without any other annotation. The map of coordinates from HXB2 reference sequence available at HIV databases was used to find the exact position of selected regions in the alignments, using the sequence as a guide (http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html). Overview of the multiple alignments was also used to identify other conserved regions.

Identification of study regions in Resistant Isolates

In addition to the mentioned regions from the conserved domain cd0548, high prevalence resistance mutations were selected from the “Mutation prevalence

According to Subtype" http://hivdb.stanford.edu/cgi-bin/MutPrevBySubtypeRx.cgi web page, from the Stanford HIV Drug Resistance Database [75, 76]. Mutations that localized within the conserved domain selected regions were preferred, but other mutations with high prevalence (>10% from the 3583 sequences used) were also chosen for analysis, if they mapped as consecutive amino acids. The following parameters were used to retrieve the data: mutation threshold >=1 % AND >=2 persons, no mixtures (to exclude possible sequencing errors), and position mutation count = one mutation per person (more than one isolate per person if different mutations are observed at the same position of the isolates). The same exact parameters were used for NRTI-experienced vs RTI naive, NNRTI-experienced (+/-NRTI) vs RTI naive, and RTI-experienced vs RTI naive.

3D Structure Analysis

Thirty crystallographic structures of HIV-1 RT available at www.rcsb.org protein databank (PDB) [77, 78] were used to perform Structural Alignments for comparison between different mutant enzymes, and wild type enzymes bound to different inhibitors. Heteroatoms were not included in the structural alignment. The following crystallographic structures were used: wild type structures 2ZD1 (1.8 Å) [79], 2RF2 (2.4 Å) [80] , 3DLE (2.5 Å) [81], 2OPP (2.55 Å) [82], 1TKT (2.6 Å) [83], 1TKZ (2.81 Å) [83], 1TKX (2.85 Å) [84], 1TL1 (2.9 Å) [83], 1TL3 (2.8 Å) [83], 1S1X (2.8 Å) [85], and mutant structures 3BGR (2.1 Å) [79], 3DLK, (1.85 Å) [86], 1S1T (2.4 Å) [86], 3DOL (2.5 Å) [81], 3DMJ (2.6 Å) [81], 1S9E (2.6 Å) [87], 1S1V (2.6 Å)

[85], 1S1W (2.7 Å) [85], 2OPQ (2.8 Å) [82], 1FKO (2.9 Å) [88], 1FKP (2.9 Å) [88], 3DOK (2.9 Å) [81], 2OPR (2.9 Å) [82], 2ZE2 (2.9 Å) [79], 1S6P (2.9 Å)[87], 1S1U (3.0 Å) [85], 2IC3 (3.0 Å) [89], 3DM2 (3.1 Å) [81]. Protein structural alignments were generated with Multiprot [90] and analyzed in PDBviewer (Deep View) [91, 92]. Mutant (MUT) and wild type (WT) structural alignments were performed independently and together. In either case 2ZD1 structure was used as reference layer. Protein structural alignments generated with MultiProt were run online, introducing PDB structure codes. Structure PDB codes were chosen to be kept in order to ensure that the first protein in the pdb file (the reference layer) was 2ZD1. A 2.5 Å threshold was chosen to generate the structural alignment with Multiprot. The comparison between WT and MUT structural alignments was performed with DeepView.

Sequence Retrieval

Curated sequences were retrieved from http://hivdb.stanford.edu/ and from http://www.hiv.lanl.gov/content/index. Resistant isolates sequences were retrieved from the Stanford HIV Resistance Database [66, 93, 94] by a therapy criterion that consisted of different combinations of three antiretrovirals, two nucleoside analogs reverse transcriptase inhibitors (NRTI) and one non-nucleoside analog reverse transcriptase inhibitor (NNRTI) respectively. Duplicates and repeated sequences were eliminated using ElimDupes http://www.hiv.lanl.gov/content/sequence/ELIMDUPES/elimdupes.html tool available at HIV databases [http://www.hiv.lanl.gov/]. Sequences were further processed to eliminate foreign

characters not compatible with FASTA format (e. p. "~"). A total of 1250 sequences of HIV-1 isolates from HIV databases curated alignment were included in the study, as well as 2263 sequences from the Stanford University HIV-1 resistance database [64-66, 93, 94]. Multiple alignment was performed with HIV align tool form HIV Databases, using Mafft [95].

In addition, two first line regimens were selected from the National Resolution No. 3442-2006 for the care of the HIV infected patient, approved for HIV infection treatment in Colombia by the Ministerio de Proteccion Social (Resolution 3442 for the Care of the HIV patient) [96] in order to retrieve sequences from resistant isolates generated during each of these drug selective pressures. The regimens are zidovudine-lamivudine-efavirenz (ZDV-3TC-EFV), which is normally the first choice, and stavudine-lamivudine-nevirapine (D4T-3TC-NVP), which is also a first line regimen.

Colombian Resistant Isolates

Twenty isolates from symptomatic HIV positive patients were included in the study, and the fasta sequences were incorporated in the group of resistant viral sequences. Only samples with informed consent, complete clinical history and viral loads over 1000 copies/ml were included. Virions from 1ml of plasma were concentrated at 15000 rpm x 99 min, resuspended in 500 ul of lysis buffer (Guanidinium thiocyanate - Biomerieux) for 10 min at room temperature. RNA extraction was performed using the semi automated Nuclisens Minimag, and Nuclisens extraction Kit, following manufacturer's instructions (Biomerieux, Marcy

l'Etoile, France) and eluted in 30 ul, of elution buffer. HIV genotyping was performed using TRUGENE HIV-1 commercial kit (Siemens Diagnostics, Deerfield, USA), and fasta sequences were further analyzed with HIV Stanford Drug Resistance Database software [97]. Quality assessment and CPG were used to evaluate probable sequencing errors and mutation prevalence respectively [75].

Multiple Alignments

Curated multiple alignments are premade alignments already edited by HIV experts and are freely available at HIV databases web site http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html. Not curated multiple alignments from resistant isolates were performed with the HIVAlign tool from HIV databases, which is an implementation of the HMMER package http://www.hiv.lanl.gov/content/sequence/HMM/HmmAlignhtml . Sequences were codon aligned with GeneCutter [http://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html], and reference sequence HXB2 was included in the alignment as guide sequence to help find selected regions. Selected regions were found using alignment editor Bio-edit and eBIOX. Four multiple alignments were used for the study, one downloaded premade-curated alignment and three generated in this study (resistant isolates). Pre-made alignment corresponds to HIV-1 Group M plus recombinants whole web 2008 sequence alignment (not resistant isolates), from HIV Databases; not pre-made alignments correspond to resistant sequences alignments.

Identification of nucleotide changes for each position of the 21 nucleotide long selected target region

A Perl script was designed to identify changes in each position of the target sequence for all the sequences aligned along each multiple alignment or the whole multiple alignment (wide and large) using a window size of 21 nucleotides. The script could find unique combinations within the selected region of the alignment and their frequencies, showing the nucleotide changes. The script was designed to include the ambiguous bases from the 21-nucleotide window interpreting each of them into all the possible combinations according to the IUPAC code and finally adding them to the frequency of its corresponding unique combination.

Design of the shRNAs that include all the variability of HIV-1 in a 21-nucleotide long target region

The same script included a Score-based program that included parameters that have been frequently found in efficient shRNAs and siRNAs and suggested by Reynolds [98] and some by Jagla et al. 2005 [99] , but in this case, the efficiency information was searched and found in the DNA sequence. That is, the program gave a score of how efficient will be the silencing induced from the RNA transcribed from the DNA sequence (template guide strand) or unique combination of the target sequence. Furthermore, the program showed which parameters were not fulfilled and which did. A score of 11 points meant that all the parameters were successfully accomplished. The script followed 6 parameters which were: wins

(window size), star_pos (position at which the script begins), end_pos (position at which script ends), threshold (tells to discard the window if the number of combinations is higher than the chosen percentage of sequences), scorethres (minimum score allowed in the window; score is based in Reynolds conditions [98], min_freq (minimum frequency allowed for each unique combination), min_freq_count (tells not to count the window if the chosen min_freq_count percentage of sequences do not meet min_freq). Percentages were selected based on the number of sequences present in each multiple alignment.

RESULTS

1. Identification of conserved regions and target sites within the conserved domain RT_Rtv

Conserved Domains and YMDD motif

Search was performed using Pol poliprotein AAB50259.1 from HXB2 reference sequence as query, and four conserved domains were found. Conserved domain RT_Rtv was identified with an E-value of 1e-85, in CDD NCBI [68, 100]. Within this domain 9 subregions were mapped that grouped DNA binding sites (green), dNTPs binding sites (dark blue), reverse transcriptase inhibitors (RTIs), binding sites (cyan blue), active sites with no other annotations (purple) and important motif YMDD (pink) [101] (Figure 1a). Table 1 shows the regions chosen for the analysis, their position in HXB2 reference genome, the important

residues within them including their position with respect to RT enzyme (HXB2 numbering system), and the nucleotide sequence for each region.

2. Resistance analysis from Colombian isolates.

A subset of 50 samples from 130 hospitalized symptomatic HIV positive patients, collected from January of 2008 to October of 2009 were analyzed in order to include the sequences from resistant isolates in the multiple alignment from resistant sequences. From those 50 sequences, only samples with viral loads over 1000 copies/milliliter were chosen for genotyping, for a total of 30. One sequence went out of study (the patient leaved the study voluntarily), 9 were repeated, and three were not recognized by COBAS AMPLICOR HIV-1 MONITOR V1.5 (ROCHE) (no viral load data for these samples) but were successfully genotyped. In total, only 25 from the 30 samples resulted in TRUGENE HIV-1 report (Siemens Diagnostics, Deerfield, USA). One sample was diagnosed as having multiple quality problems (e. i. stop codons - frame shifts – too many unusual residues), the sequencing run was revised and complete assay repetition was not enough to achieve a higher quality and was not included (Sup Table 1). Thirty six percent of the protease (PR) sequences had any mutation, 12% of the RT sequences had any protease inhibitor (PI) mutation, 24% had any nucleoside reverse transcriptase inhibitor (NRTI) mutation, 16% had any NNRTI mutation, 12% had any of both types of mutations. Four percent of the PRRT sequences had any NRTI, NNRTI and PI 3-class resistance mutations [102]. PI mutations seen in the data set were I54T and M64I, with a set prevalence of 4% and 8%. Major NRTI thymidine analog

mutations (TAMs) were D67N (4%), M184V (16%), L210W (4%), and T215Y (4%), major non-thymidine analog mutations were K65R (8%) and K70R (4%), and F77L (4%) was the only multi-NRTI resistance mutation seen in the data set. Major NNRTI mutations were K103N (8%), Y181C (4%), Y188L (4%), and P225H (4%). Only one mutation G152R, indicative of APOBEC3G –mediated editing was seen, in only one sequence. No more hyper mutated sites were identified, and at least one atypical mutation was observed for each of the sequences analyzed. Identity of PR sequences with HXB2 reference sequence was of 95.16%, for RT of 94.29%, and for the complete sequences PRRT of 78.68% (mean value). Five samples showed high level resistance to NRTIs, four of these against 2 NRTI and one against 3 NRTIs. No high level resistance against PIs was observed for the studied set of samples (Sup Table 2).

3. Resistance mutations within selected regions

Resistant mutations localized near or within the selected regions were identified in the corresponding alignments. Table 2 shows the wild type residues, their position with respect to reference sequence HXB2, the mutated residue and prevalence based on HIV Drug Resistance Database data. Regions are named based on table 1. Our hypothesis is based on that if mutations are selected by certain regimen they could become “fixated” under selective drug pressure, and in the fact that if they match in mutable positions exactly within conserved active site residues, this could increase the “conservation” of the whole site, making it a possible target for gene therapy. From the regions that have mutations within or next to them, the most appropriate to target are those which

have the same mutation selected by both types of antiretrovirals (NRTIs – nucleoside RT inhibitors, NNRTIs - non nucleoside RT inhibitors), and even better if there is only one observed mutation for the residue. Following this criterion, chosen regions were number 4, 7, and 8. Since the targeted nucleotide region is about 21 nts long (7 triplets), regions 4 and 7 must include an extra residue at any end. Nonetheless, the other regions were screened and in case of region 9, since the residues next to each end of the initially chosen residues are not conserved and are mutable under drug selective pressure, windows including both residues were analyzed.

4. 3D structure analysis

Structural alignment showed variability between mutant and wild type RT enzymes. Thirty crystallographic structures were used for the alignments. Structures had a mean residue number of 942.55, which ranged from 914 to 979 residues. Structure Resolution ranged form 1.8 Å to 3.1 Å. Structure 2ZD1, was used as reference layer since it has the higher resolution 1.8 Å, and corresponds to a wild type enzyme. Figure 1b shows the structural alignment of wild type RT enzymes, with 774 aligned residues and 0.81 root mean square deviation (rms). Structural alignment evidenced high variability between non-mutant RTs. Being structures of the same wild type enzyme from the same virus we expected a better alignment, also that structures were not so variable between them and thus a bigger number of residues participating in the alignment. Figure 1c corresponds to the structural alignment of resistant (mutant) enzymes that showed even higher variability, with only 590 residues

participating in the alignment and 0.88 rms. This variability is evident in the superimposed alignments showed in Figure 1d. Structures have different mutations, and are also in complex with different ligands. Variability is evident in the number of residues involved in the alignment, which is far less than the number used for wild type RT structure alignment.

Even though structural alignments were performed with Deep View and Multiprot, WT and MUT structural alignment corresponds to the best generated with Multiprot (best from three).

5. Multiple Alignments

A total of 2533 sequences were retrieved from the Stanford drug resistance database, 286 were duplicates, and thus were eliminated. After editing characters, multiple alignment was generated with HIV align tool from HIV Databases tools (http://www.hiv.lanl.gov/content/sequence/HMM/HmmAlign.html) which is an implementation of HMMER package. Sequences were aligned by HMM (Hidden Markov Models) model HIV-1/SICcpz, allowing 5 codons to compensate for frame shift [103]. In addition, sequences selected under the regimens ZDV-3TC-EFV and D4T-3TC-NVP were also retrieved from this database.

Sequences for the premade curated alignment were retrieved form the HIV Databases, and consisted of curated sequences updated to 2008 and correspond to Group M HIV-1, including recombinant forms. This multiple alignment was not further edited.

6. shRNAs guide strands that include the variability found.

Curated Alignment

Along the curated multiple alignment from Group M +Recombinants 2008 (HIV databases), only 2 regions arose a combination of Scores and number of unique combinations (frequency) that could be useful for gene therapy (e. i. 1 and 7). Different conditions of the script parameters were tested and found that there was no region without, at least one unique combination with a score of 4. A higher threshold for minimum score rendered no results. From all the different conditions evaluated, the following were finally used to test the selected regions from curated alignment: threshold 20%, Score-threshold = 4, min_fre_count 10%, min_freq = 6. Results had to be exhaustively examined because not all the windows thrown were useful.

Table 3 depicts the results for region 3230- 3243, windows 3234 and 3237, and for region 3648 – 3662, window 3648. In all the alignment, there were no other regions with less than 46 unique combinations in which the most frequent ones rendered minimum scores of 7, the minimum score required for silencing.

Region 3230-3243, window 3234, with 46 total unique combinations, had scores between 4 and 9, but the most frequent unique combination occurred 959 times, with a score of 8. The next most frequent unique combinations occurred 40, 32, 29 and 24 times in the alignment, and only one had a score under 7, the one that occurred 40 times. Together without taking into account this last combination, the group of sequences accounted for the 89.3 % of the 1256 sequences in the alignment. All the remaining combinations had frequencies under 10, mostly of 1 (this is also true for the other regions and

windows obtained). About the score conditions, none of the most frequent combinations met conditions 1, 6 and 8, that correspond to having an A in position 3, and having an A or T in positions 15 and 17.

Window 3237, with 54 unique combinations showed better results, being only between 8 and 9, having 9 for the most frequent combination, which occurs 942 times. The next most frequent unique combinations occurred 40, 32, 27, and 23 times, all of them with a score of 8. All together these combinations accounted for the 87.6 % of the sequences in the alignment. In this case, score conditions 1 and 11, having A in position 3 and having A or T in position 20, were the only conditions not met by all the 5 most frequent unique combinations. Individually, combinations with a score of 8 had one extra condition which they did not meet.

Window 3548, from region 7, with 46 unique combinations, had the best scores, being between 9 and 10, but had twice the number of most frequent unique combinations with respect to the other windows. The frequency of the most common unique combinations was more equivalent between each other, being 167, 161, 27 and 25 for scores of 10, and 255, 165, 113, 91, 25 and 22, for scores of 9. All together correspond to the 86.6 % of the sequences.

Other three windows from region 3391 – 3420 gave a number of unique combinations under 61, and scores of 7, 8, and 9 but included a gap. The same was true for region 3628- 3662, window 3648, which scored over 8. Gap editing did not improve the results probably because nucleotide positions were changed, changing the whole score.

Resistant sequences

The script threshold conditions used for analyzing the curated alignment were not useful for analyzing the resistant sequences multiple alignment. Using these same conditions the analysis rendered too many windows. The only condition kept was score threshold at 4. In fact conditions had to be adequated to each alignment since each one has a different number of sequences and in turn representative variability. For each regimen alignment it was possible to find target regions with fine scores, specifically in the same regions in which it was possible to find targets for the curated alignment. We find the highest score of 11, and also of 10, but rarely frequently. The small number of combinations in target sequences for shRNA design found for DAT-3TC-NVP seems to be a matter of number of sequences present in the alignment. Nonetheless the number of sequences in the ZDV-3TC-EFV is close to the one of the curated alignment, and gave more combinations than the alignment of the other regimen, even when the curated alignment has sequences from all group M plus recombinants (Table 3).

Curated alignment is supposed to be more divergent than regimen sequence alignments, which are principally, but not uniquely composed of B subtype strains. In the case of these alignments, silencing would be determined by the number and selection of the target sequences that can be introduced into a gene therapy vector. For D4T-3TC-NVP the number of combinations accounted for 78.02% (window 179), 87.9% (window 264), 83.52% (window 459), 73% (window 565) and for ZDV-3TC-EFV alignment the number of combinations accounted for 89.07% (window 72) of the variability present in the

alignment, and 88.06% for window 73. Results for each window may seem different, but only *in vitro* studies will open light in the true silencing efficiency.

7. BLAST

Blastn searches were performed from CLC DNA Workbench Desktop for each of the unique combinations in each region, against Human genomic plus transcripts (test/gpipe/9606) all contig_and_rna NCBI database (table 4), and against Human subset of ESTs database (table 5). Search was limited by Entrez Organism *Homo sapiens*, using low complexity and mask for lower case as filters. In addition, default values for word size, match, mismatch, gap open cost, and gap extension cost, were used. Only the minus strand hits found are shown. In all cases Hits mapped to unique sequence tagged sites (in human chromosomes, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, X genomic contigs, reference or alternate assembly), in no case an overlap of 100% was seen. Most of the hits repeated along the region, but most E-values were not significant (1.2 -4.5). Only most significant E-values of 0.08 and 0.3 are shown (Tables 4 and 5). Hits on minus strands are important because expressed sequences from these hits could be targeted. From the significant E-values (0.3), 11 targeted minus strands, and from the two most significant E-values (0.08), both targeted minus strands. In the case of targets from resistant sequences, number of hits ranged from 5 to 100 for each combination, each having at least one significant hit with significant E-value MISMA TABLA. Only important ESTs hits are shown for AZV-3TC-EFV, like IL-2, white blood cells, B cells etc., since many of the hits were ESTs from different types of cancer in

different tissues not infected by HIV. Important hits are those expressed by gene therapy target cells or progenitors.

DISCUSSION

This is the first approach to a novel design of shRNAs guide strands, based on the scored search of a group of sequences that target most or all the subset of wild-type and drug-resistant RT variants. We developed a script in order to follow the established parameters of effective guide strand for shRNAs from Reynolds et al.2004 [98]. In our approach shRNAs guide strand sequences are directed to silence resistant variants selected during antiretroviral therapy, avoiding therapy failure and consequences that result from it. A set of guide strands was also find for HIV-1 Group M plus recombinant HIV-1 viruses.

shRNAs design is difficult due to the multiple requirements to achieve an efficient *in vivo* silencing and to all the parameters that had to be carefully followed. Available programs are normally directed to siRNA and not shRNA design [104-106], and some authors have shown that these are not useful to correctly predict the silencing efficiency of shRNAs, perhaps because processing occurs through different cellular pathways [107-109]. Developed tools like Gene Link (http://www.genelink.com/sirna/shRNAi.asp) do not allow the use of more than one aligned sequence, which is a necessary issue when trying to design silencing molecules against error-prone viruses like HIV. Recently miRanda [110] was used to check *in silico,* target sites for miRNAs in the 3' UTR of the HIV-1 molecular clone LAI genome, which were then tested with rna22 [111, 112]. This approach was based on using miRNAs and shRNAs with partial complementarities instead of

a total identity, which was thought to be absolutely necessary for shRNAs efficient silencing. Because of this premise, there was no need for a multiple alignment but it surely would have been informative.

The search for highly conserved regions in HIV-1 genome has been done for several RNAi screening, and has been based on multiple alignments and other types of sequence comparisons [45, 51, 56]. Nonetheless, no approach has analyzed in detail the conservation based on protein function, which is thought to be the least variable. In multiple alignments of HIV nucleotide sequences the observed variability is a result of the number of sequences present in the alignment, unless the whole set of sequences analyzed are representative of the virus variability which will be more than a million sequences for a conserved region based on the different number of DNA codons for each specific amino acid. In the particular case of HIV-1, multiple alignments from HAART treated isolates as well as from untreated isolates reveal enormous variability even on very conserved regions. As it was evident in the multiple alignments from this study, variability in the most restricted or conserved regions is still high enough to make any design strategy based on finding a consensus shRNA guide strand sequence inefficient in terms of long term or indefinite silencing, especially in a human HIV-1 infected patient. More over, even though multiple shRNAs targeted to conserved regions of the viral genome have proven to inhibit HIV-1 replication, viral escape has shown to be unimpeded [113-117]. Hypothetically, in an infected patient all the possible variants of each of the targeted conserved sequences must also be targeted in order to achieve life-long silencing. Eventually, this would increase the number of shRNAs that would have to be expressed from a single or multiple vectors, which

may be a limitation if the number and size exceeds certain limit [118]. The targeting of multiple conserved regions remains a constant strategy since it has showed encouraging results [43, 45, 51, 53, 54, 56, 118, 119], even though one should ask if silencing efficiency is the result of targeting one same sequence twice or more (times the number as conserved sequences are being targeted), due to targeting of more number of viral variants once with the only conserved sequence guide strand from all to which complementarity occurred, or both.3

Still there is another question: if all escape variants are indeed the result of "resistance" to RNAi therapy or if they are naturally occurring variants which are not being targeted and instead are being selected.

To counteract this problem, we decided to target a unique highly conserved region that tested with our bioinformatics tool showed the lowest number of different forms or unique combinations along the multiple alignments, that met a score of at least 7 for the most frequent ones, and that this most frequent variants were very abundant so that all together represented the 80% or more of the sequences present in the alignments. Conserved regions were chosen based on a protein-function approach, aiming to find a region with near seven consecutive conserved amino acids. Seven is the minimum score required for silencing [98]. Unfortunately, the highest score (11) was not frequent, and even worse, it was mainly found in low conserved regions (data not shown). Consequently shRNAs requirements were easily accomplished in inappropriate regions rather than in the conserved ones. It must be outlined that though a certain sequence has a very low frequency, e. g. one, it may have been the dominant variant from an infected patient. This means that the unknown intrapatient variability may modulate

behavior of virus resistant quasispecies turning the least frequent variant into the dominant one. The fact is that conservation depended on the number of DNA codons that could result in a particular amino acid sequence, that is why in the YMMD motif the conservation was less than the expected. Curiously, regions 1 and 7 participate in DNA and dNTP binding, with only 2 and 4 non-consecutive amino acids respectively. These a.a. are involved in the two tasks mentioned, and had the best scores with the least number of combinations, under 48, from which two of them occurred in more than 900 sequences. Together, using the group of combinations of two or more regions will help fighting against viral variants. These results were obtained with the alignment of whole group M plus recombinants, however, nucleotide variation observed in subtypes different from B is important because this variability is acceptable in terms of enzyme function. This means that irrespective of the subtype the mutation can happen, and under selective drug pressure has a higher probability of appearance, without meaning that this variant will convert to another subtype. Still, there is a difference between naive viral variants and resistant variants, which not only has to be addressed *in vitro*, but which was the reason we analyzed sequences from resistant variants. Since resistant variants under certain drug inhibitor share resistance mutations, without changing drug selective pressure these mutations will become selected and thus conserved in the resistant viral population. Based on this fact, we searched within the previously selected conserved regions, resistant mutations that could increase the number of conserved consecutive amino acids in order to test it for shRNAs guide strand design. In fact we found much more variability, for example for regimen ZDV-3TC-EFV which was composed of 1381 sequences, the most

frequent combination occurred 825 times. This is less than in the curated alignment of whole group M plus recombinants, which is much more divergent. These results evidence the variability capacity of HIV virus, which undoubtedly is an obstacle for HIV gene therapy in an infected patient. Thus, depending on the regimen, the conserved nucleotide position in multiple alignments for YMDD ranged from 1 to 8 of the 12 positions (other regimen specific alignments were used, data not shown). Unfortunately, this range is in certain degree a matter of the number of sequences used in the alignment, which means we have not arrived to the minimum number of sequences that represent HIV variability. For the region with the least number of unique combinations, amino acid reference sequence HXB2 is WPLTEEK, which can be formed by 512 different nucleotide sequences. It could be less if there is a preference for codon usage in HIV virus. Looking at the reference alignment for pol polyprotein (HIV Databases) the following mutations were seen: W24R, P25LTS, T26SA, E27KAGR, E28K, K29ER, which results in 286,654,464 possibilities. Since a region of 21 nt corresponds to 7 amino acids, the reason why the region of the YMMD motif being the most conserved did not show as expected, the least number of unique combinations, lies in the nearby amino acids, which can be coded by several DNA codons. Other conserved regions with nearby amino acids codified by less number of DNA codons could render better results.

Recently, Mcintyre et al. 2009 analyzed a big set of sequences and find 96 shRNAs for maximal coverage of the virus genome [56]. Even if our work had a similar approach, the final aim is different since they chose one shRNA for region and not the most frequent variants for each of them, which will reduce variability, as we did. For each of the regions they found, there are surely other frequent

variants that should be taken into account. If not, having more regions will result in increased variability reducing the therapy effectiveness, by means of the selection of different naturally occurring variants plus the resistant ones. On the other hand, targeting accessory genes could help in reducing pathogenicity of the clinical infection, but not necessarily viral replication, even though some accessory genes have a role in essential steps of viral cycle [18, 20, 21, 120-122]. Since nef protein is important in the context of the infection, as well as other accessory genes, this could explain the efficiency of silencing HIV in a humanized mice model, perhaps they are needed for viral fitness [47].

Our target was specifically RT because it cannot be deleted from the genome, besides it is the main source of virus variability. Furthermore, it is one of the targets of HAART, and there is a lot of information about mutations that occur under selective drug pressure [19, 62, 69, 123, 124]. Reverse transcription is absolutely necessary for viral integration and infection of other cells, so its silencing would reduce not only variability, but also the number of cells that become actively and latently infected, and thus HIV cell reservoirs [18, 125]. We did not target other essential enzymes [122], because targeting various enzymes or regions simultaneously would add variability (as was explained before), increasing the number of shRNAs needed for an efficient silencing. Nonetheless, multiple conserved regions could be the better approach as it has been previously proposed [43, 51, 56, 119], but only if the most frequent combinations for each conserved region are used for a shRNAs combinatorial design, in order to be able to target the most frequent variants. They will not necessarily be targeted by all the

groups of shRNAs simultaneously, but targeting a viral variant by at least one shRNA could result in silencing.

It might be helpful to study sequences from different specific regimens in order to find the better subset of shRNAs that should silence the emergent resistant isolates. This would be preferable to changing the previous regimen after a HAART failure, which will only change the viral population that is being selected, increasing the probability of failure with each regimen change. On the other hand, it might be useful to evaluate other regions with our approach in order to find better Score values with less number of shRNAs molecules, and perhaps find along the viral genome a region with a set of sequences that can cover the greatest percentage of virus variants, like for example LTR, which was proved to be very conserved [56]. Full-length genome sequencing of resistant viruses would be convenient for detecting target regions either in *pol* or in other genes, which can also target sequences from therapy-naive patients.

Efficient silencing was achieved *in vivo* in a mouse model using shRNA targeting a region within *nef* gene [47] even when before it was seen in vitro that nef gene was deleted as a route of escape [116]. In the pre-made curated multiple alignments we found no conservation in any of the 21-nucleotide positions used for the nef sequence (data not shown). In other words, silencing was directed to certain variants of the virus. In humans, quasispecies may be an obstacle for silencing, and using shRNAs against few variants of a region, as an initial therapy, will select the untargeted variants or induce emergence of resistance. The same is true in highly “conserved regions” of HIV-1 that surprisingly, reveal unexpected variability, which was even more evident in resistant isolates, and structure

alignments. Since mainly all the structures used were in complex with a specific ligand, different to all of them, observed variability may be caused by flexibility of the enzyme or structural conformational motions in order to achieve binding. In addition, variability is also a result of individual polymorphisms' that do not count as resistant mutations, but help in conserving the whole structure. Curiously, the conserved regions under study showed less structural alignment than we expected. Probably, even without mutations within this study region, polymorphisms distant from important residues create smooth structural conformational changes that are advantageous and maintain functional activity undamaged, as has being proposed before [126, 127]. So based on amino acid variability of conserved regions observed in structural alignments, we expected nucleotide variability to be high as we found.

Sequences identified here, might result in very efficient silencing once incorporated to an efficient hairpin [128], depending on intrapatient variability. Difference in silencing will only be evident proving the sequences *in vitro* and *in vivo*. If the number of sequences introduced in the expression vector does not become a problem, sequences from region 2 could work well since they have the higher scores. On the other hand, having a less number of molecules targeting more number of variants could be better in terms of expression [43]. The fact is that the combination of sequences chosen for the shRNA design is crucial in HIV-1 gene therapy for a life-long efficacy. We do not discard the possibility of finding even better sequences for shRNA design. For this, the script must be improved, and other parameters should be included.

Most Blast hits resulted in non-significant E-values probably because short sequences have higher probability of occurring in the database. Nonetheless, even if the E-values we obtained are generally not considered as significant it is necessary to be careful since the targets exist in the databases, and in most cases identity was 100%. It was favorable that most of the probable sequences or unique combinations did not had a 100% overlapping with blast hits, only 9 from ZDV-3TC-NVP had full 21 nt alignment but not identity. In this case overlapping is important since the size of 21 nts is necessary for the intracellular-silencing pathway, and because normally shRNA are designed to target a fully complementary sequence [129, 130]. This year Liu et al., proposed a partially complementary approach for HIV inhibition, using shRNAs and artificial miRNAs, in which targeting not perfect sites by these two types of molecules resulted in silencing: miRNA and shRNA were capable of inhibiting multiple imperfect targets in the 3' UTR (untranslated region) of the viral genome. Our ESTs blast results showed that the overlapping hit aligned usually in the center, in some cases completely at the 3' end or at the 5' end of the guide strand. In these cases, identity was 100%. Hit sequences were usually shorter, but they ranged from 16 to 21 nts, which could represent high-risk off-target effects over the human mRNA sequences, particularly if they are expressed by the hematopoietic cell line. On the other hand, mutations in viral sequences were seen in different positions with no apparent pattern. All the possible targets of a partially complementarity strategy should be analyzed against ESTs databases in order to be sure to reduce off-target effects, since partially complementary sequences open the possibility of targeting cell genes that other wise would not be targeted if complete

complementarity was required for silencing. In addition, the blast results, obtained with the blast search against Human EST databases should be also addressed *in vitro* in order to confirm that there is no silencing of these human mRNAs species, even though overlapping is fewer than 95%. Each entry should be carefully analyzed in order to identify the specific cells that express the mRNA, when in development is it expressed and if it is already silenced by natural miRNAs. Many of the blast hits were ESTs associated with different types of cancer, perhaps in these cases off-target effects could be advantageous (e.g. Burritt's Lymphoma).

Also, using a multiple alignment of only a specific subtype, e. p. subtype B, could bring better results, since it should be more conserved than the reference multiple alignment used in this study that includes sequences from whole group M including recombinants. Our results are encouraging towards a more exhaustive study of HIV variability, and towards the finding of much better shRNAs based on subtype specific multiple alignments, which could be useful for a region based attempt. The fact that RT-Rtv domain is evolutionarily conserved among kingdoms, explains why it is relatively variable in HIV-1. Based on this argument, the next step would be to target unique viral sequences like *tat* or *rev* genes in combination with conserved regions of *env* (gp 120), targeting most frequent variants, in order to reduce viral variability for immune system, and help it control the infection. A possible region could be the residues important for CXCR5 binding, which are hidden until CD4 receptor binding with gp 120 [131].

In another hypothetical approach shRNAs could be used as a complement to HAART, where shRNAs should be directed against the specific resistant variants that emerge during a specified regimen. In this case, while HAART inhibits the

dominant and highly fit variants, shRNAs could inhibit a high percentage of the less fit emergent resistant variants. Perhaps specific mutations could be induced voluntarily in order to select variants that can be easily targeted by gene therapy. Our approach resulted in template guide strand sequences that can be used for shRNA design in order to be tested alone or under the specified regimen selective pressure. In a future prospect, unless we can find a better small number of combinations of sequences that could target all the variants, the ideal situation should be to initiate both therapies. shRNAs should only be started soon after virologic failure is evident, since most mutations occur under drug selective pressure. Indeed, further studies are needed in order to find regimens whose resistance pattern can be specified and better controlled using the least number of shRNAs. Proving real *in vivo* efficiency in a combination therapy as well as identifying off-target and non-specific off-target effects of the shRNAs molecules is absolutely required before starting any approach.

ACKNOWLEDGEMENTS

This study was sponsored by Instituto Nacional de Salud (INS) in Bogotá D.C., Colombia, and Departamento Administrativo de Ciencia Tecnología e Innovación COLCIENCIAS Bogotá D.C., Colombia. Authors thank Dr. Andrés Paez Martinez at Instituto Nacional de Salud in Bogotá, for his competent assistance in the preparation of this manuscript.

DECLARATION OF CONFLICT OF INTERESTS

None

FINANCIAL SUPPORT

Departamento Administrativo de Ciencia Tecnología e Innovación COLCIENCIAS, Bogota D.C., Colombia.

SUPPLEMENTARY MATERIAL

Table 1. Clinical data for genotyped samples

No. Samples	Recluse	55
	Analyzed	25
	Genotyped	25
Gender	Males	20
	Females	5
Age	Range	23 - 65
	Mean	30
Viral Load copies/ml	Range	1040 - >750000
	Mean	290,890
Final Condition	Alive	20
	Dead	5

1

Table 2. Antiretroviral Resistance data

Class	Inhibitor	Resistance (number of sequences)			
		Potential low level	Low level	Intermediate	High level
NRTI	3TC	1	0	1	4
	ABC	3	0	2	1
	AZT	0	0	1	1
	D4T	1	1	1	1
	DDI	1	0	3	
	FTC	1	0	1	4
	TDF	0	0	3	
NNRTI	DLV	0	0	1	3
	EFV	0	0	0	4
	ETR	1	2	1	0
	NVP	0	0	0	4
PRI	ATV/r	3	0	0	0
	DRV/r	0	0	0	0
	FPV/r	3	0	0	0
	IDV/r	3	0	0	0
	LPV/r	3	0	0	0
	NFV	0	2	1	0
	SQV/r	0	2	0	0

TPV/r	2	0	0	0

3TC lamivudine, ABC abacabir, AZT zidovudine, D4T stavudine, DDI didanosine, FTC emtricitabine, TDF tenofovir, DLV delavirdine, EFV efavirenz, ETR etravirine, NVP nevirapine, ATV/r atazanavir, DRV/r darunavir, FPV/r fosamprenavir, IDV/r indinavir, LPV/r lopinavir, NFV nelfinavir, SQV saquinavir, TPV tripanavir. Antiretrovirals ARV/r, ritonavir boosted.

REFERENCES

1. Cohen, J (2000). Searching for the epidemic's origins. *Science* **288**: 2164-2165.

1. Cohen, J (2000). Searching for the epidemic's origins. *Science* **288**: 2164-2165.

2. Cohen, MS (2006). Amplified transmission of HIV-1: missing link in the HIV pandemic. *Trans Am Clin Climatol Assoc* **117**: 213-224; discussion 225.

3. Mazumder, A, Engelman, A, Craigie, R, Fesen, M, and Pommier, Y (1994). Intermolecular disintegration and intramolecular strand transfer activities of wild-type and mutant HIV-1 integrase. *Nucleic Acids Res* **22**: 1037-1043.

4. Wu, Y (2004). HIV-1 gene expression: lessons from provirus and non-integrated DNA. *Retrovirology* **1**: 13

5. Chiu, TK, and Davies, DR (2004). Structure and function of HIV-1 integrase. *Curr Top Med Chem* **4**: 965-977.

6. Garber, DA, Silvestri, G, and Feinberg, MB (2004). Prospects for an AIDS vaccine: three big questions, no easy answers. *Lancet Infect Dis* **4**: 397-413.

7. Tripathi, P, and Agrawal, S (2007). Immunobiology of human imunodeficiency virus infection. *Indian J Med Microbiol* **25**: 311-322.

8. Potter, SJ, Chew, CB, Steain, M, Dwyer, DE, and Saksena, NK (2004). Obstacles to successful antiretroviral treatment of HIV-1 infection: problems & perspectives. *Indian J Med Res* **119**: 217-237.

9. Sethi, AK (2004). Adherence and HIV drug resistance. *HIV Clin Trials* **5**: 112-115.

10. Bennett, DE, Bertagnolio, S, Sutherland, D, and Gilks, CF (2008). The World Health Organization's global strategy for prevention and assessment of HIV drug resistance. *Antivir Ther* **13 Suppl 2**: 1-13.

11. Boasso, A, and Shearer, GM (2008). Chronic innate immune activation as a cause of HIV-1 immunopathogenesis. *Clin Immunol* **126**: 235-242.

12. Moriuchi, M, and Moriuchi, H (2002). In vitro reactivation of HIV-1 by stimulation with herpes simplex virus infection. *Sex Transm Dis* **29**: 308-309.

13. Weinberg, MS, and Morris, KV (2006). Are viral-encoded microRNAs mediating latent HIV-1 infection? *DNA Cell Biol* **25**: 223-231.

14. King, SR, Duggal, NK, Ndongmo, CB, Pacut, C, and Telesnitsky, A (2008). Pseudodiploid genome organization AIDS full-length human immunodeficiency virus type 1 DNA synthesis. *J Virol* **82**: 2376-2384.

15. Cornelissen, M, van den Burg, R, Zorgdrager, F, Lukashov, V, and Goudsmit, J (1997). pol gene diversity of five human immunodeficiency virus type 1 subtypes: evidence for naturally occurring mutations that contribute to drug resistance, limited recombination patterns, and common ancestry for subtypes B and D. *J Virol* **71**: 6348-6358.

16. van der Kuyl, AC, and Cornelissen, M (2007). Identifying HIV-1 dual infections. *Retrovirology* **4**: 67.

17. Greatorex, J (2004). The retroviral RNA dimer linkage: different structures may reflect different roles. *Retrovirology* **1**: 22.

18. Freed, EO (2001). HIV-1 replication. *Somat Cell Mol Genet* **26**: 13-33.

19. Basu, VP, Song, M, Gao, L, Rigby, ST, Hanson, MN, and Bambara, RA (2008). Strand transfer events during HIV-1 reverse transcription. *Virus Res* **134**: 19-38.

20. Basavapathruni, A, and Anderson, KS (2007). Reverse transcription of the HIV-1 pandemic. *FASEB J* **21**: 3795-3808.

21. Cochrane, AW, McNally, MT, and Mouland, AJ (2006). The retrovirus RNA trafficking granule: from birth to maturity. *Retrovirology* **3**: 18.

22. Martinez-Picado, J, and Martinez, MA (2008). HIV-1 reverse transcriptase inhibitor resistance mutations and fitness: a view from the clinic and ex vivo. *Virus Res* **134**: 104-123.

23. Alimonti, JB, Ball, TB, and Fowke, KR (2003). Mechanisms of CD4+ T lymphocyte cell death in human immunodeficiency virus infection and AIDS. *J Gen Virol* **84**: 1649-1661.

24. Collins, KR, *et al.* (2002). Human immunodeficiency virus type 1 (HIV-1) quasispecies at the sites of Mycobacterium tuberculosis infection contribute to systemic HIV-1 heterogeneity. *J Virol* **76**: 1697-1706.

25. Quan, Y, Liang, C, Brenner, BG, and Wainberg, MA (2009). Multidrug-Resistant Variants of HIV Type 1 (HIV-1) Can Exist in Cells as Defective Quasispecies and Be Rescued by Superinfection with Other Defective HIV-1 Variants. *J Infect Dis* **200**: 1479-1483.

26. Quinones-Mateu, ME, and Arts, EJ (2002). Fitness of drug resistant HIV-1: methodology and clinical implications. *Drug Resist Updat* **5**: 224-233.

27. Nigam, PK, and Kerketta, M (2006). AIDS vaccine: present status and future challenges. *Indian J Dermatol Venereol Leprol* **72**: 8-18.

28. Heeney, JL, Dalgleish, AG, and Weiss, RA (2006). Origins of HIV and the evolution of resistance to AIDS. *Science* **313**: 462-466.

29. Holmes, EC (2001). On the origin and evolution of the human immunodeficiency virus (HIV). *Biol Rev Camb Philos Soc* **76**: 239-254.

30. Piacenti, FJ (2006). An update and review of antiretroviral therapy. *Pharmacotherapy* **26**: 1111-1133.

31. Cavalcanti, AM, Lacerda, HR, Brito, AM, Pereira, S, Medeiros, D, and Oliveira, S (2007). Antiretroviral resistance in individuals presenting therapeutic failure and subtypes of the human immunodeficiency virus type 1 in the Northeast Region of Brazil. *Mem Inst Oswaldo Cruz* **102**: 785-792.

32. Kanzaki, LI, Ornelas, SS, and Arganaraz, ER (2008). RNA interference and HIV-1 infection. *Rev Med Virol* **18**: 5-18.

33. Marathe, JG, and Wooley, DP (2007). Is gene therapy a good therapeutic approach for HIV-positive patients? *Genet Vaccines Ther* **5**: 5.

34. Agrawal, N, Dasaradhi, PV, Mohmmed, A, Malhotra, P, Bhatnagar, RK, and Mukherjee, SK (2003). RNA interference: biology, mechanism, and applications. *Microbiol Mol Biol Rev* **67**: 657-685.

35. Shafer, RW, Rhee, SY, and Bennett, DE (2008). Consensus drug resistance mutations for epidemiological surveillance: basic principles and potential controversies. *Antivir Ther* **13 Suppl 2**: 59-68.

36. Lucas, GM (2005). Treatment of experienced patients and antiretroviral resistance issues from the 12th CROI. *Hopkins HIV Rep* **17**: 1-3, 12.

37. Bhavan, KP, Kampalath, VN, and Overton, ET (2008). The aging of the HIV epidemic. *Curr HIV/AIDS Rep* **5**: 150-158.

38. Leung, RK, and Whittaker, PA (2005). RNA interference: from gene silencing to gene-specific therapeutics. *Pharmacol Ther* **107**: 222-239.

39. Tijsterman, M, Ketting, RF, and Plasterk, RH (2002). The genetics of RNA silencing. *Annu Rev Genet* **36**: 489-519.

40. Saayman, S, Barichievy, S, Capovilla, A, Morris, KV, Arbuthnot, P, and Weinberg, MS (2008). The efficacy of generating three independent anti-HIV-1 siRNAs from a single U6 RNA Pol III-expressed long hairpin RNA. *PLoS One* **3**: e2602.

41. Scherer, L, Rossi, JJ, and Weinberg, MS (2007). Progress and prospects: RNA-based therapies for treatment of HIV infection. *Gene Ther* **14**: 1057-1064.

42. Sano, M, Li, H, Nakanishi, M, and Rossi, JJ (2008). Expression of long anti-HIV-1 hairpin RNAs for the generation of multiple siRNAs: advantages and limitations. *Mol Ther* **16**: 170-177.

43. Liu, YP, *et al.* (2009). Combinatorial RNAi Against HIV-1 Using Extended Short Hairpin RNAs. *Mol Ther*.

44. Miyano-Kurosaki, N, and Takaku, H (2006). Gene silencing of virus replication by RNA interference. *Handb Exp Pharmacol*: 151-171.

45. ter Brake, O, Konstantinova, P, Ceylan, M, and Berkhout, B (2006). Silencing of HIV-1 with RNA interference: a multiple shRNA approach. *Mol Ther* **14**: 883-892.

46. Kirchhoff, F (2008). Silencing HIV-1 In Vivo. *Cell* **134**: 566-568.

47. ter Brake, O, *et al.* (2009). Evaluation of safety and efficacy of RNAi against HIV-1 in the human immune system (Rag-2(-/-)gammac(-/-)) mouse model. *Gene Ther* **16**: 148-153.

48. Touraine, JL, Sanhadji, K, and Sembeil, R (2003). Gene therapy for human immunodeficiency virus infection in the humanized SCID mouse model. *Isr Med Assoc J* **5**: 863-867.

49. Shacklett, BL (2008). Can the new humanized mouse model give HIV research a boost. *PLoS Med* **5**: e13.

50. von Eije, KJ, and Berkhout, B (2009). RNA-interference-based gene therapy approaches to HIV type-1 treatment: tackling the hurdles from bench to bedside. *Antivir Chem Chemother* **19**: 221-233.

51. von Eije, KJ, ter Brake, O, and Berkhout, B (2008). Human immunodeficiency virus type 1 escape is restricted when conserved genome sequences are targeted by RNA interference. *J Virol* **82**: 2895-2903.

52. Tsygankov, AY (2009). Current developments in anti-HIV/AIDS gene therapy. *Curr Opin Investig Drugs* **10**: 137-149.

53. ter Brake, O, t Hooft, K, Liu, YP, Centlivre, M, von Eije, KJ, and Berkhout, B (2008). Lentiviral vector design for multiple shRNA expression and durable HIV-1 inhibition. *Mol Ther* **16**: 557-564.

54. von Eije, KJ, ter Brake, O, and Berkhout, B (2009). Stringent testing identifies highly potent and escape-proof anti-HIV short hairpin RNAs. *J Gene Med* **11**: 459-467.

55. Naito, Y, *et al.* (2007). Optimal design and validation of antiviral siRNA for targeting HIV-1. *Retrovirology* **4**: 80.

56. McIntyre, GJ, Groneman, JL, Yu, YH, Jaramillo, A, Shen, S, and Applegate, TL (2009). 96 shRNAs designed for maximal coverage of HIV-1 variants. *Retrovirology* **6**: 55.

57. Joseph, A, Sango, K, and Goldstein, H (2009). Novel mouse models for understanding HIV-1 pathogenesis. *Methods Mol Biol* **485**: 311-327.

58. Koyanagi, Y, Tanaka, Y, Ito, M, and Yamamoto, N (2008). Humanized mice for human retrovirus infection. *Curr Top Microbiol Immunol* **324**: 133-148.

59. Zhang, L, Kovalev, GI, and Su, L (2007). HIV-1 infection and pathogenesis in a novel humanized mouse model. *Blood* **109**: 2978-2981.

60. Rodriguez-Arenas, MA, *et al.* (2006). Delay in the initiation of HAART, poorer virological response, and higher mortality among HIV-infected injecting drug users in Spain. *AIDS Res Hum Retroviruses* **22**: 715-723.

61. Richman, DD, *et al.* (2004). The prevalence of antiretroviral drug resistance in the United States. *AIDS* **18**: 1393-1401.

62. Baldwin, C, and Berkhout, B (2007). HIV-1 drug-resistance and drug-dependence. *Retrovirology* **4**: 78.

63. Berkhout, B (2009). Toward a durable anti-HIV gene therapy based on RNA interference. *Ann N Y Acad Sci* **1175**: 3-14.

64. Shafer, RW (2006). Rationale and uses of a public HIV drug-resistance database. *J Infect Dis* **194 Suppl 1**: S51-58.

65. Rhee, SY, Gonzales, MJ, Kantor, R, Betts, BJ, Ravela, J, and Shafer, RW (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res* **31**: 298-303.

66. Shafer, RW, Jung, DR, Betts, BJ, Xi, Y, and Gonzales, MJ (2000). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res* **28**: 346-348.

67. Ding, J, Hughes, SH, and Arnold, E (1997). Protein-nucleic acid interactions and DNA conformation in a complex of human immunodeficiency virus type 1 reverse transcriptase with a double-stranded DNA template-primer. *Biopolymers* **44**: 125-138.

68. Marchler-Bauer, A, *et al.* (2007). CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* **35**: D237-240.

69. Back, NK, *et al.* (1996). Reduced replication of 3TC-resistant HIV-1 variants in primary cells due to a processivity defect of the reverse transcriptase enzyme. *EMBO J* **15**: 4040-4049.

70. Coggill, P, Finn, RD, and Bateman, A (2008). Identifying protein domains with the Pfam database. *Curr Protoc Bioinformatics* **Chapter 2**: Unit 2 5.

71. Finn, R, Griffiths-Jones, S, and Bateman, A (2003). Identifying protein domains with the Pfam database. *Curr Protoc Bioinformatics* **Chapter 2**: Unit 2 5.

72. Finn, RD, *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res* **36**: D281-288.

73. Hunter, S, *et al.* (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**: D211-215.

74. Mulder, N, and Apweiler, R (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* **396**: 59-70.

75. Gifford, RJ, *et al.* (2009). The calibrated population resistance tool: standardized genotypic estimation of transmitted HIV-1 drug resistance. *Bioinformatics* **25**: 1197-1198.

76. Rhee, SY, *et al.* (2006). HIV-1 pol mutation frequency by subtype and treatment experience: extension of the HIVseq program to seven non-B subtypes. *AIDS* **20**: 643-651.

77. Berman, HM, *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res* **28**: 235-242.

78. Berman, H, Henrick, K, and Nakamura, H (2003). Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**: 980.

79. Das, K, *et al.* (2008). High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: strategic flexibility explains potency against resistance mutations. *Proc Natl Acad Sci U S A* **105**: 1466-1471.

80. Zhao, Z, *et al.* (2008). Novel indole-3-sulfonamides as potent HIV non-nucleoside reverse transcriptase inhibitors (NNRTIs). *Bioorg Med Chem Lett* **18**: 554-559.

81. Ren, J, *et al.* (2008). Structural basis for the improved drug resistance profile of new generation benzophenone non-nucleoside HIV-1 reverse transcriptase inhibitors. *J Med Chem* **51**: 5000-5008.

82. Ren, J, *et al.* (2007). Relationship of potency and resilience to drug resistant mutations for GW420867X revealed by crystal structures of inhibitor complexes for wild-type, Leu100Ile, Lys101Glu, and Tyr188Cys mutant HIV-1 reverse transcriptases. *J Med Chem* **50**: 2301-2309.

83. Hopkins, AL, *et al.* (2004). Design of non-nucleoside inhibitors of HIV-1 reverse transcriptase with improved drug resistance properties. 1. *J Med Chem* **47**: 5912-5922.

84. Freeman, GA, *et al.* (2004). Design of non-nucleoside inhibitors of HIV-1 reverse transcriptase with improved drug resistance properties. 2. *J Med Chem* **47**: 5923-5936.

85. Ren, J, Nichols, CE, Chamberlain, PP, Weaver, KL, Short, SA, and Stammers, DK (2004). Crystal structures of HIV-1 reverse transcriptases mutated at codons 100, 106 and 108 and mechanisms of resistance to non-nucleoside inhibitors. *J Mol Biol* **336**: 569-578.

86. Bauman, JD, *et al.* (2008). Crystal engineering of HIV-1 reverse transcriptase for structure-based drug design. *Nucleic Acids Res* **36**: 5083-5092.

87. Das, K, *et al.* (2004). Roles of conformational and positional adaptability in structure-based design of TMC125-R165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant HIV-1 variants. *J Med Chem* **47**: 2550-2560.

88. Ren, J, Milton, J, Weaver, KL, Short, SA, Stuart, DI, and Stammers, DK (2000). Structural basis for the resilience of efavirenz (DMP-266) to drug resistance mutations in HIV-1 reverse transcriptase. *Structure* **8**: 1089-1094.

89. Das, K, Sarafianos, SG, Clark, AD, Jr., Boyer, PL, Hughes, SH, and Arnold, E (2007). Crystal structures of clinically relevant Lys103Asn/Tyr181Cys double mutant HIV-1 reverse transcriptase in complexes with ATP and non-nucleoside inhibitor HBY 097. *J Mol Biol* **365**: 77-89.

90. Shatsky, M, Nussinov, R, and Wolfson, HJ (2004). A method for simultaneous alignment of multiple protein structures. *Proteins* **56**: 143-156.

91. Guex, N, and Peitsch, MC (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**: 2714-2723.

92. Schwede, T, Kopp, J, Guex, N, and Peitsch, MC (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* **31**: 3381-3385.

93. Shafer, RW, Stevenson, D, and Chan, B (1999). Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database. *Nucleic Acids Res* **27**: 348-352.

94. Kantor, R, Machekano, R, Gonzales, MJ, Dupnik, K, Schapiro, JM, and Shafer, RW (2001). Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database: an expanded data model integrating natural language text and sequence analysis programs. *Nucleic Acids Res* **29**: 296-299.

95. Katoh, K, Misawa, K, Kuma, K, and Miyata, T (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059-3066.

96. Granados, CAD, Álvarez, C, Prada, G, Martínez, FL, and Sarmiento, CA (2006). Guía para el manejo de VIH/sida Basada en la evidencia In *Modelo de gestión programática en VIH/sida*. Programa de Apoyo a la Reforma de Salud Ajuste a los Planes de Beneficio y la Unidad de Pago por Capitación. MPS, COLOMBIA.

97. Liu, TF, and Shafer, RW (2006). Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin Infect Dis* **42**: 1608-1618.

98. Reynolds, A, Leake, D, Boese, Q, Scaringe, S, Marshall, WS, and Khvorova, A (2004). Rational siRNA design for RNA interference. *Nat Biotechnol* **22**: 326-330.

99. Jagla, B, *et al.* (2005). Sequence characteristics of functional siRNAs. *RNA* **11**: 864-872.

100. Marchler-Bauer, A, *et al.* (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* **37**: D205-210.

101. Huang, H, Chopra, R, Verdine, GL, and Harrison, SC (1998). Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* **282**: 1669-1675.

102. Prajapati, DG, Ramajayam, R, Yadav, MR, and Giridhar, R (2009). The search for potent, small molecule NNRTIs: A review. *Bioorg Med Chem* **17**: 5744-5762.

103. Eddy, SR (1998). Profile hidden Markov models. *Bioinformatics* **14**: 755-763.

104. Gong, W, *et al.* (2006). Integrated siRNA design based on surveying of features associated with high RNAi effectiveness. *BMC Bioinformatics* **7**: 516.

105. Ren, Y, Gong, W, Zhou, H, Wang, Y, Xiao, F, and Li, T (2009). siRecords: a database of mammalian RNAi experiments and efficacies. *Nucleic Acids Res* **37**: D146-149.

106. Tilesi, F, Fradiani, P, Socci, V, Willems, D, and Ascenzioni, F (2009). Design and validation of siRNAs and shRNAs. *Curr Opin Mol Ther* **11**: 156-164.

107. Bartel, DP (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**: 281-297.

108. Zhang, C (2008). MicroRNomics: a newly emerging approach for disease biology. *Physiol Genomics* **33**: 139-147.

109. Zhang, B, Wang, Q, and Pan, X (2007). MicroRNAs and their regulatory roles in animals and plants. *J Cell Physiol* **210**: 279-289.

110. John, B, Enright, AJ, Aravin, A, Tuschl, T, Sander, C, and Marks, DS (2004). Human MicroRNA targets. *PLoS Biol* **2**: e363.

111. Miranda, KC, *et al.* (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**: 1203-1217.

112. Liu, YP, Gruber, J, Haasnoot, J, Konstantinova, P, and Berkhout, B (2009). RNAi-mediated inhibition of HIV-1 by targeting partially complementary viral sequences. *Nucleic Acids Res*.

113. Boden, D, Pusch, O, and Ramratnam, B (2007). Overcoming HIV-1 resistance to RNA interference. *Front Biosci* **12**: 3104-3116.

114. Boden, D, Pusch, O, Lee, F, Tucker, L, and Ramratnam, B (2003). Human immunodeficiency virus type 1 escape from RNA interference. *J Virol* **77**: 11531-11535.

115. Boden, D, Pusch, O, and Ramratnam, B (2004). HIV-1-specific RNA interference. *Curr Opin Mol Ther* **6**: 373-380.

116. Das, AT, *et al.* (2004). Human immunodeficiency virus type 1 escapes from RNA interference-mediated inhibition. *J Virol* **78**: 2601-2605.

117. Westerhout, EM, Ooms, M, Vink, M, Das, AT, and Berkhout, B (2005). HIV-1 can escape from RNA interference by evolving an alternative structure in its RNA genome. *Nucleic Acids Res* **33**: 796-804.

118. Liu, YP, Haasnoot, J, and Berkhout, B (2007). Design of extended short hairpin RNAs for HIV-1 inhibition. *Nucleic Acids Res* **35**: 5683-5693.

119. Liu, YP, Haasnoot, J, ter Brake, O, Berkhout, B, and Konstantinova, P (2008). Inhibition of HIV-1 by multiple siRNAs expressed from a single microRNA polycistron. *Nucleic Acids Res* **36**: 2811-2824.

120. Bennasser, Y, and Jeang, KT (2006). HIV-1 Tat interaction with Dicer: requirement for RNA. *Retrovirology* **3**: 95.

121. Costin, JM (2007). Cytopathic mechanisms of HIV-1. *Virol J* **4**: 100.

122. Klimas, N, Koneru, AO, and Fletcher, MA (2008). Overview of HIV. *Psychosom Med* **70**: 523-530.

123. Bakhouch, K, *et al.* (2009). The prevalence of resistance-associated mutations to protease and reverse transcriptase inhibitors in treatment-naive (HIV1)-infected individuals in Casablanca, Morocco. *J Infect Dev Ctries* **3**: 380-391.

124. Boyle, BA, *et al.* (2008). Update on antiretroviral therapy: the 15th CROI. *AIDS Read* **18**: 273-278, C273.

125. Brenner, B, *et al.* (2004). Persistence of multidrug-resistant HIV-1 in primary infection leading to superinfection. *AIDS* **18**: 1653-1660.

126. Gutteridge, A, and Thornton, J (2004). Conformational change in substrate binding, catalysis and product release: an open and shut case? *FEBS Lett* **567**: 67-73.

127. Tousignant, A, and Pelletier, JN (2004). Protein motions promote catalysis. *Chem Biol* **11**: 1037-1042.

128. Ding, H, Liao, G, Wang, H, and Zhou, Y (2007). Asymmetrically designed siRNAs and shRNAs enhance the strand specificity and efficacy in RNAi. *J RNAi Gene Silencing* **4**: 269-280.

129. Rao, DD, Vorhies, JS, Senzer, N, and Nemunaitis, J (2009). siRNA vs. shRNA: similarities and differences. *Adv Drug Deliv Rev* **61**: 746-759.

130. Cohen, MS, Hellmann, N, Levy, JA, DeCock, K, and Lange, J (2008). The spread, treatment, and prevention of HIV-1: evolution of a global pandemic. *J Clin Invest* **118**: 1244-1254.

Figure 1.

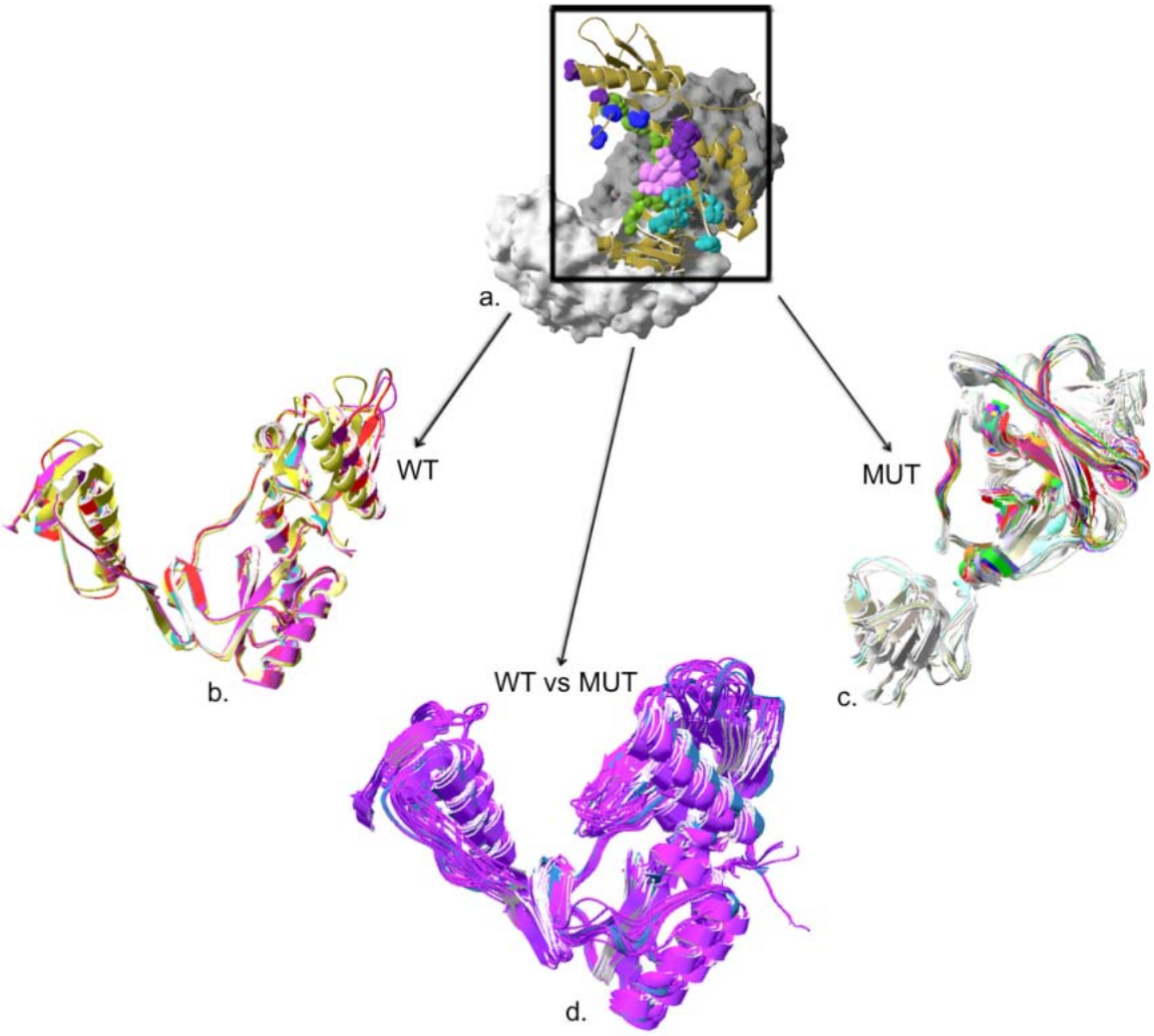

Fig1. a. RT crystallographic structure 2ZD1 (1.8 Å) highlighting inside the squared frame the residues within selected regions. Dark gray p66 subunit, in light gray p55, in dark blue active site residues involved dNTP binding (K65, R72, D110, V111, G112, D113, A114, Y115, Q151), in green active site residues involved in DNA Binding (L74, V75, D76, R78, N81, E89, Q91, L92, I94, G152, K154, P157, M230, G231), in purple active site residues with no specific annotations (W24, P25, F61), in pink YMDD motif (Y183, M184, D185, D186), and in light blue residues involved in NNRTI binding (L100, K101, K102, K103, V179, Y188, G190, F227; not conserved). Ribbon shows continuity between amino-acid chains. b. Active site structural alignment for 10 different RT wild type structures, in ribbon (RMS 0.81). Bound ligands and heteroatoms (hidden) were not taken into account for the structural alignment. Subunit p51 (A chain) and residues 297 to 545, are hidden, but aligned. Reference layer: structure 2ZD1. b. Structural alignment for 19 different RT mutant enzymes, in ribbon. Selected regions are shown in highlighted colors, other not important residues in grayscale. Only 590 residues were structurally aligned. Structural alignment was generated with Multiprot [90].

1

Table 1. Conserved residues and selected regions chosen for the analysis, their position in HXB2 reference genome, the important residues within them including their position with respect to RT enzyme (HXB2 numbering system), and the nucleotide sequence for each region.

Region	Residue No.	Amino acid	Function	DNA codon in HXB2 genome	Evaluated region*
1	24	W	DNA Binding and Active Site	2619	$_{2616}$caaTGGCAAttgaca$_{2624}$
	25	P		2622	
2	61	F	Active Site	2730	$_{2814}$TTTgccataaagAAA$_{2831}$

	65	K	dNTP Binding and Active Site	2742	
3	72	R	dNTP Binding and Active Site	2763	$_{2763}$AGAaaaTTAGTAGATttcAGAgaacttAAT$_{2792}$
	74	L	DNA	2769	
	75	V	Binding	2772	
	76	D	and	2775	
	78	R	Active	2781	
	81	N	Site	2790	

4	89 91 92 94	E Q L I	DNA Binding and Active Site	2814 2820 2823 2829	$_{2814}$GAAgttCAATTAggaATA$_{2831}$
5	100 101 102 103	L K K K	NNRTI Binding site	2847 2850 2853 2856	$_{2847}$TTAAAAAAGAAA$_{2859}$
6	110 111 112 113 114 115	D V G D A Y	dNTP Binding and Active Site	2877 2880 2883 2886 2889 2892	$_{2877}$GATGTGGGTGATGCATAT$_{2894}$

7	151	Q	dNTP Binding and Active Site	3000	$_{3000}$CAGGGAtggAAAggatcaCCA$_{3020}$
	152	G	DNA	3003	
	154	K	Binding and	3009	
	157	P	Active Site	3018	
8	179	V	NNRTI Binding site	3084	$_{3084}$GTTatctactatcaaTACAGTGATGATttgTATgtaGGA$_{3119}$
	183	Y	NNRTI Binding	3096	

		Site, DNA Binding and Active Site	
184	M	variable	3099
		dNTP Binding	
185	D	and	3102
		Active Site	
186	D	Active Site	3105
188	Y	NNRTI	3111

	190	G	Binding site NNRTI Binding site	3117	
9	227	F	NNRTI Binding site	3228	$_{3228}$TTCctttggATGGGTtatgaactc$_{3251}$
	230	M	DNA Binding and Active Site	3237	
	231	G		3240	

*DNA codons in capslock correspond to the residues in the region. Nucleotide sequences correspond to the HXB2 reference genome, positions according to the HXB2 numbering system (map of coordinates). Each

nucleotide position will be considered the starting position of a window from which the following upstream 20 nts are included

1 **Table 2.**

Table 2. Prevalence of Resistant Mutations that reside inside selected regions

	HXB2		Mutations Prevalence		
Region[1]	**Position**	**Residue(s)**	**NRTI[2]**	**NNRTI[3]**	**RTI[4]**
2	60	V	V (12)	V (16)	V (14)
	61	F^	-	-	-
	62	A	V (2.2)	V (4.8)	V (14)
	63	I	-	-	-
	64	K	R (1.6)	R (2.1), H (1.2)	R (1.9)
	65	K^	R (1.3)	R (2.6)	R (2.1)
	66	K	-	-	-
	67	D	N (26), G (1.1)	K (2.6), G (3.3)	N (38), G (2.5)
4	89	E^	0		
	90	V	I (3.7)	I (2.4)	I (3.3)
	91-94	Q^, L^, G, I^	-	-	-

7	151	Q^	M (4.0)	M (2.3)	M (3.4)
	152-157	G^, W, K^, G, S, P^	-	-	-
8	178	I	M (7.8), L (6.7)	M (7.4), L (6.2)	M (7.5), L (6.4)
	179	V	I (3.3)	I (9.6), D (1.7)	I (7.0), D (1.3)
	180	I	-	-	-
	181	Y	-	C (23), I (1.1)	C (14)
	182	Q, Y^	-	-	-
	184	M	V (49)	V (50), I (1.9)	V (50), I (1.4)
	185	D^,D^, L	-	-	-
	188	Y	-	L (5.9)	L (3.5)
	189	V	-	I (2.3)	I (1.5)
	190	G	-	A (16), S (2.9)	
9	227	F^	-	L (2.5)	L (1.6)
	228	L	-	H (17), R (5.9)	H (12), R (4.2)
	229	W	-	-	-

230	M^	-	-	L (1.8)
231	G^	-	-	-

[1]Corresponds to the numeration given to selected regions from table 1.

[2]NRTI prevalence was calculated from 3583 sequences exposed to these type of drugs

[3]NNRTI prevalence was calculated form 5070 sequences exposed to these type of drugs

[4]RTI prevalence was calculated from 17167 sequences exposed to either type of these drugs

[2,3,4]Mutations prevalence data is available at HIV Drug Resistance Database

^Residues directly involved in enzyme activity

In shade, selected regions.

1
2

Table 3.

Table 3. Scores, unique combinations (guide strand template) and sequences for the regions selected for silencing

Alignment	Region of study	Region in HXB2	Window	No. of U.C[2]	U.C[2]	Sequence	Score	Frequence	Score Conditions[3]													
									1	2	3	4	5	6	7	8	9	10	11	12	13	
Curated Alignment Group M plus Recombinants	1	2616 - 2630	3234	46	1	AGGAGCAGATGATACAGTATT	8	959	N	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y	Y	
					2	AGGAGCAGATGATACAGTGTT	5	40	N	Y	N	N	Y	N	Y	N	Y	N	Y	Y	Y	
					3	AGGAGCAGATGACACAGTATT	8	32	N	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y	Y	
					4	AGGGGCAGATGATACAGTATT	8	29	N	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y	Y	
					5	AGGAGCAGATGATACAGTACT	7	24	N	Y	Y	Y	Y	N	Y	N	Y	Y	N	Y	Y	
		2616 - 2630	3237	54	1	AGCAGATGATACAGTATTAGA	9	942	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	
					2	AGCAGATGATACAGTGTTAGA	8	40	N	Y	Y	Y	Y	Y	N	Y	Y	Y	N	Y	Y	
					3	AGCAGATGACACAGTATTAGA	8	32	N	N	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	
					4	GGCAGATGATACAGTATTAGA	9	27	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	
					5	AGCAGATGATACAGTACTAGA	8	23	N	Y	Y	Y	Y	Y	Y	N	Y	Y	N	Y	Y	
	7	3000 - 3014	3648	47	1	CCAGTATTTGCCATAAAAAAG	10	167	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
					2	CCAATATTTGCCATAAAAAAG	10	161	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
					3	CCAGTATTTGCTATAAAAAAG	10	27	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
					4	CCAATATTTGCTATAAAAAAG	10	25	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
					5	CCAGTATTTGCCATAAAGAAA	9	255	Y	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	
					6	CCAGTATTTGCTATAAAGAAA	9	165	Y	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	
					7	CCAATATTTGCCATAAAGAAA	9	113	Y	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	
					8	CCAATATTTGCTATAAAGAAA	9	91	Y	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	
					9	CCAGTGTTTGCCATAAAGAAA	9	25	Y	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	
					10	CCAATATTTGCAATAAAGAAA	9	22	Y	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	

D4T-3TC-NVP (stavudine, lamivudine, nevirapine) 91 sequences	2	179	28	1	TATTTGCTATAAAGAAAAAGA	8	14	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y
				2	TATTTGCTATAAAGAAAAAAG	9	8	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y
				3	TATTTGCCATAAAGAAAAAAG	9	20	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y
				4	TATTTGCTATAAAGAAAAAGG	8	16	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y
				5	TATTTGCCATAAAGAAAAAGG	8	7	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y
				6	TATTTGCCATAAAAAAGAAGG	7	6	N	Y	Y	Y	Y	Y	Y	N	Y	Y	N	N	Y
	4	264	11	1	GGAAGTTCAATTAGGAATACC	8	60	Y	N	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y
				2	GGAAGTTCAGTTAGGAATACC	8	7	Y	N	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y
				3	GGAAATTCAATTAGGAATACC	7	7	Y	N	Y	Y	Y	N	Y	Y	Y	Y	N	N	Y
				4	GGAGGTTCAATTAGGAATACC	8	6	Y	N	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y
	7	459	11	1	GAAAGGATCACCAGCAATATT	8	54	Y	N	Y	Y	Y	N	Y	Y	Y	Y	Y	N	Y
				2	GAAAGGATCACCGGCAATATT	7	14	Y	N	Y	Y	N	N	Y	Y	Y	Y	Y	N	Y
				3	GAAGGGATCACCAGCAATATT	9	8	Y	N	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y
	8	565	21	1	GTAGGATCTGACTTAGAAATA	9	35	Y	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y
				2	GTAGGATCTGATTTAGAAATA	9	18	Y	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y
				3	GTAGGATCTGATTTAGAAATA	9	14	Y	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y
ZDV-3TC-EFV (zidovudine. Lamivudine, efavirenz) 1381 sequences	1	72	74	1	GCCATTGACAGAAGAAAAAAC	9	5	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
				2	GCCATTGACAGAAGAAAAAAT	9	825	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
				3	GCCACTGACAGAAGAAAAAAT	9	56	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
				4	GCCATTGACAGAAGAAAAGAT	6	21	N	N	N	N	Y	Y	Y	Y	Y	N	Y	Y	Y
				5	GCCCTTGACAGAAGAAAAAAT	9	20	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
				6	GCCATTGACAGAGGAAAAAAT	8	42	N	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
				7	GCCATTAACAGAAGAGAAAAT	8	15	N	N	Y	Y	Y	Y	N		Y	Y	Y	Y	Y
				8	GCCATTGACGGAAGAAAAAAT	9	28	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
				9	GCCATTGTCAGAAGAAAAAAT	9	13	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
				10	GCCATTGACAAAAGAAAAAAT	8	11	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y
				11	GCCATTGACAGAAGAGAAAAT	8	144	N	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y
				12	GCCATTAACAGAAGAAAAAAT	9	33	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
				13	GCCATTGACAAAAGAGAAAAT	7	17	N	N	Y	Y	Y	Y	N	Y	Y	Y	Y	N	Y
		73	77	1	CCATTGACAGAAGAGAAAATT	8	31	Y	N	Y	Y	N	N	Y	Y	Y	Y	Y	Y	Y
				2	CCATTGTCAGAAGAAAAAATA	9	13	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y

3	CCATTGACAGAAGAAAAGATA	8	21	Y	N	Y	Y	N	Y	Y	Y	N	Y	Y	Y	Y
4	CCATTGACAGAAGAGAAAATA	8	114	Y	N	Y	Y	N	N	Y	Y	Y	Y	Y	Y	Y
5	CCATTGACAAAAGAGAAAATA	7	17	Y	N	Y	Y	N	N	Y	Y	Y	Y	Y	N	Y
6	CCATTGACAGAGGAAAAAATA	9	42	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
7	CCATTGACAGAAGAAAAAATT	9	13	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
8	CCATTGACGGAAGAAAAAATA	9	28	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
9	CCCTTGACAGAAGAAAAATA	8	20	N	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
10	CCATTAACAGAAGAAAAAATA	9	33	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
11	CCATTGACAGAAGAAAAAATA	9	813	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
12	CCATTAACAGAAGAGAAAATT	8	15	Y	N	Y	Y	N	N	Y	Y	Y	Y	Y	Y	Y
13	CCACTGACAGAAGAAAAAATA	9	56	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y

[1]M.A. Multiple Alignment

[2]U.C.= Unique Combinations (with respect the total No. of sequences in the M.A)

[3]Score conditions:

1. Has A in position 3
2. Has T in position 10
3. No C/G in position 19, but has A
4. No C/G in position 19
5. No G in postion 13
6. Has A or T in position 15
7. Has A or T in position 16
8. Has A or T in position 17
9. Has A or T in position 18
10. Has A or T in position 19
11. Has A or T in position 20
12. No more than three consecutive A's from position 1 to 14
13. The sum of C or G is less than 11

Y = accomplished

N= not accomplished

TABLE 4. Most significant Blast Hits against UniSTS only hits in minus strands are shown

Alignment	No.	HIT	U.C	E-value	Bit score	Overlap	Hit frame/strand
Curated alignment	1	NT_005403	R1,w2,c4	0.3	36.18	76.19	Minus
	1	NT_005403	R1,w2,c5				
	2	NW_921618					
	3	NT_016297	R2,w1,c3	0,3	36,18	76,19	Minus
	4	NW_922073					
	5	NT_025741	R2,w1,c6	**0,08**	38,16	80,95	Minus
	6	NW_923184					
	7	NT_009714	R2,w1,c8	0,3	36,18	76,19	Minus
	8	NT_016354					
	9	NW_925328					
	10	NW_922162					
	11	NT_025741	R2,w1,c9	0,3	36,18	76,19	Minus
	12	NW_923184					
	13	NT_022517	R2,w1,c10	0,3	36,18	76,19	Minus
	14	NT_004487					
	15	NW_921651					
	16	NW_926128					

Blast search against UniSTS database was only perform with Curated Alignment combinations.

1

Table 5. Blast Most significant Hits against Human ESTs database

Alignment	Region	HIT	U.C	E-value	Bit score	Overlap
Curated	7	BF240717	R7, 8	0,21231	36,1753	85,71
		DA099283	R7, 4 R7, 8	0,838917	34,1929	
		AW503859	R7, 4 R7, 8	0,838917	34,1929	80,95
		D17012	R7, 4 R7, 8	0,838917	34,1929	
D4T-3TC-NVP (stavudine-lamivudine-nevirapine)	7	BF240717	R7, 8	0,21231	36,1753	85,71
		DA099283	R7, 4 R7, 8	0,838917	34,1929	
		AW503859	R7, 4 R7, 8	0,838917	34,1929	80,95
		D17012	R7, 4 R7, 8	0,838917	34,1929	
ZDV-3TC-EFV (zidovudine-lamivudine-efavirenz)[a]	1	DA020938	R1,72-3	3,44E-03	32,21	90,48
			R1,72-2	0,84	34,19	90,48
			R1,73-13	0,01	40,14	85,71
		T34054	R1 (72-5, 73-9)	0,84	34,19	71,43
		DB095531	R1 (72-6, 73-6)	0,21	36,18	76,19
		BF380718		0,84	34,19	71,43
		BX325384				
		BX340654				
		AL581553				
		AL560565	R1 (72-1, 72-2, 73-7, 73-11)	0,21	36,18	76,19
		AL558996				
		BX325200				
		BU633649				
		BQ019860				

	AV757747	72-12	0,84	34,19	71,43
	BF184320	73-7			90,48

[a]Only important hits with significant scores are shown