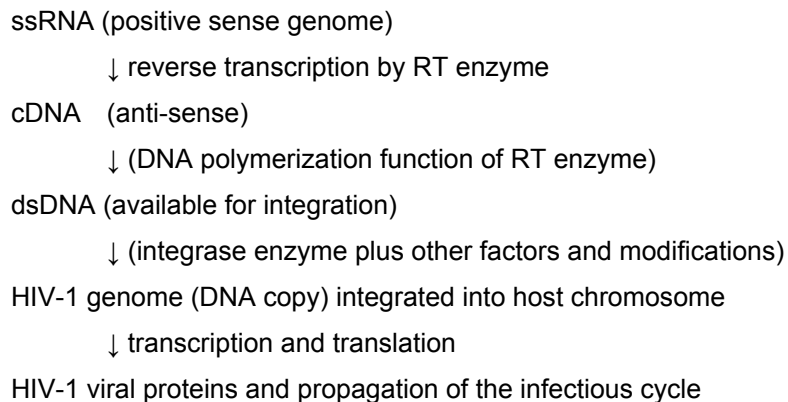


Data Mining for the Clinical Laboratorian Part 1: An HIV-1 Integrase Model

Dawn R. Maghakian M.S., MP (ASCP), CLSp (MB)

The human genome databases are clearly a wealth of information, but using bioinformatics tools to find the most relevant *in-silico* information can be a daunting task amongst the vast amount of information available. However, as molecular-based methods continue to advance in sophistication and complexity, and we produce more high-complexity test results affecting clinical decision making, it has never been more important for the Clinical Scientist to gain a functional understanding of the database resources available in molecular diagnostics. Importantly, this type of data mining information may be used for quality control of sequence-based test results. In this first review of the series, the integrase protein coding sequence (CDS) of the HIV-1 genome is used as a model to demonstrate how to obtain focused bioinformatics of interest in Clinical Lab Science.

The Human Immunodeficiency Virus Type 1 (HIV-1), (family Retroviridae, genus Lenitvirus) has a linear single-stranded ribonucleic acid (ssRNA) genome which is 9181 nucleotides in length and of a positive sense. [1, 2] Each virion contains two copies of its genome and encodes for 9 genes. [2, 3] During the infectious cycle the virus further produces a cDNA anti-sense genomic copy by utilizing reverse transcriptase (RT). [4]



This cDNA may serve as a template for polymerization of the dsDNA viral genome copy utilizing the DNA polymerase function of the viral RT enzyme. The dsDNA copy of the HIV-1 genome can then be prepared for integration into the host cell genome by a strand displacement and ligation. [4] FDA-approved anti-HIV-1 therapy may be directed at multiple targets designed to disrupt the ability of the virus to carry out this replication and integration cycle. [4] The HIV-1 integrase enzyme is part of a large nucleoprotein complex which is introduced into the CD4+ helper T cell host and transcribed from the HIV-1 core during the infectious process. [4, 5] This complex also includes two HIV-1 RNA transcripts, and the RT enzyme. [4, 5]

The integrase enzyme has become a target for therapy, especially when standard protease and reverse transcriptase inhibitor treatment drug resistance occurs. [4, 6–8] The enzyme functions to facilitate integration of the dsDNA HIV-1 genome into the host cell; this function is mandatory for the further transcription of viral proteins from the host chromosome. [4] INI (integrase inhibitor) drugs are expected to be used as primary therapy in combination with PIs (protease inhibitors) and RTIs (reverse transcriptase inhibitors) after completion of clinical trials to treat naïve individuals currently in progress. [6–8] For the purpose of this model we will look at the location of three common integrase resistance mutations, Q148 (changed to H, K, or R), N155H, and E92Q which are reported to account for a great majority of documented treatment failures. [6–8]

To begin our inquiry we will navigate to the NCBI home page <http://www.ncbi.nlm.nih.gov/> and select the “Genome” database option from the search function drop-down menu. [9] After typing in the term “HIV-1” into the adjacent text field and hitting “Go” you will be presented with a link to

Figure 1: HIV-1 Integrase coding sequence (CDS).

NC_001802.1 gi: 9629357 (Accessed 3-23-2008) <http://www.ncbi.nlm.nih.gov/>
(start and stop codon highlighted in blue)

```
3721 tggaggaaat gaacaagtag ataaattagt cagtgtgga atcaggaaag tacta ttt
3781 agatggaata gataaggccc aagatgaaca tgagaaatat cacagtaatt ggagagcaat
3841 ggctagtgat ttaacctgc cacctgtagt agcaaaagaa atagtagcca gctgtgataa
3901 atgtcagcta aaaggagaag ccatgcatgg acaagtagac tgtagtccag gaatatggca
3961 actagattgt acacatttag aaggaaaagt tatcctggta gcagttcatg tagccagtgg
4021 atatatagaa gcagaagtta ttccagcaga aacaggggcag gaaacagcat attttcttt
4081 aaaattagca ggaagatggc cagtaaaaac aatacatact gacaatggca gcaatttcac
4141 cggtgctacg gttagggccg cctgttggtg ggcgggaatc aagcaggaat ttggaattcc
4201 ctacaatccc caaagtcaag gagtagtaga atctatgaat aaagaattaa agaaaattat
4261 aggacaggta agagatcagg ctgaacatct taagacagca gtacaaatgg cagtattcat
4321 ccacaatttt aaaagaaaag gggggattgg ggggtacagt gcaggggaaa gaatagtaga
4381 cataatagca acagacatac aaactaaaga attacaaaaa caaattaca aaattcaaaa
4441 ttttcgggtt tattacaggg acagcagaaa tccactttgg aaaggaccag caagctcct
4501 ctggaagggt gaagggggcag tagtaataca agataatagt gacataaaag tagtgccaag
4561 aagaaaagca aagatcatta gggattatgg aaaacagatg gcaggtgatg attgtgtggc
4621 aagtagacag gatgag gat agaacatgga aaagtttagt aaaacaccat atgtatgtt
```

the Human immunodeficiency virus 1, complete genome listed as NC_001802. This alphanumeric accession number denotes the NCBI curated reference sequence (RefSeq) for this organism. All RefSeq accession numbers consist of two characters followed by an underscore and six numbers. [9] After clicking this hyperlink you will be presented with the HIV-1 complete genome data page.

To continue, locate the hyperlink NC_001802 in the “Genome info” column of the data table at the top of this page. This will open the Entrez Nucleotide report. On this page you have a choice of viewing FEATURES. For instance you can select the hyper link CDS to view only the FASTA coding sequence (CDS) that is translated into protein. For this exercise we will view the entire HIV-1 genome sequence that is displayed at the bottom of this initial report. The integrase nucleotide sequence begins at nucleotide 3776 [Fig. 1]. Also listed under the features of this page are nine (9) links to the genes that are part of the HIV-1 genome; selecting any of these links will bring you more specific information as desired. A hyperlink feature was also available

to access the gene-specific information from the initial HIV-1 complete genome data page on the data table. By scrolling down in the FEATURES of this current page you are on, you will find the integrase peptide is part of the gag-pol mature peptide sequence (mat_peptide) NP_705928.1. The integrase protein (p32) is a component of the polyprotein encoded by the gag-pol gene locus or Gene ID 155348. [1, 2] The integrase mat_peptide FASTA sequence is listed as starting and ending at nucleotides 3776 to 4639. [Fig. 1] Selecting on this hyperlink to NP_705928.1 brings up the FASTA amino acid sequence at the bottom of the page. [Fig. 2]

Figure 2: HIV-1 Integrase Amino Acid Sequence.

NP_705928.1 gi: 25121908 (Accessed 3-23-2008) <http://www.ncbi.nlm.nih.gov/>

ORIGIN

```
1  ftdgidkaqd enekyhsnwr amasdnipp vvakeivasc dkcqikgeam ngqvdcspgi
61  wqidctnleg kvilvavhva sgyieaevip aetgqetayt iiklagrwpv ktintangsn
121 ftgatvraac wwagikqetg ipynpqsqgv vesmnkeikk iigqvrdaqe niktavqmav
181 fihntkrkkg iggysageri vdiatdiqt kelqkqtki qntrvyrds rnpwkgpak
241 llwkgegavv iqdndsikvv prrkakiird ygkqmagddc vasrqded
```

By integrating the information you have just found you can create a table of alignment which will help to define the location of the three integrase resistance mutations, Q148 (H, K, or R), N155H, and E92Q. [Table 1] The FASTA sequences may also be used to relate your sequence alignment to other databases. For HIV two such resources are the Stanford University HIV Drug Resistance Database <http://hivdb.stanford.edu/pages/algs/HIVdb.html> [10] and the Los Alamos National Laboratory (Department of Health and Human Services and The National Institutes of Health) HIV Sequence Database <http://www.hiv.lanl.gov/>. At either of these on-line resources you will find a wealth of options for further learning.

As clinical laboratory scientists, we often are called upon to answer questions regarding results we produce. Increasingly in molecular diagnostics these results can be quite complex. By independently confirming the sequence of mutation in question, we increase our effectiveness to provide accurate data that may aid physicians in discriminating therapy for a difficult to treat disease.

REFERENCES

1. Scosyrev, E. (2006) An overview of the human immunodeficiency virus featuring laboratory testing for drug resistance. *Clin Lab Sci.* 19. (4) 231 – 245.
2. National Center for Biotechnology Information, Ref Seq. NC_001802, accessed 3-10-2008.
3. The Los Alamos National Laboratory (Department of Health and Human Services and The National Institutes of Health) HIV Sequence Database <http://www.hiv.lanl.gov/>.
4. Pommier, Y. et al. (2005) Integrase Inhibitors to treat HIV/AIDS. *Nat Rev Drug Dis.* 4. 236 – 248.
5. Craige, R. (2001) HIV Integrase, a brief overview from chemistry to therapeutics. *JBC.* 276 (26). 23213 – 23216.
6. MERCK & CO. Inc. (2007) Package insert. ISENTRESS™ (raltegravir) tablets.
7. Collins, S. (2007) Conference Reports: Integrase inhibitors and resistance. HIV Treatment Bulletin. 8(6/7). Published by i-Base <http://www.i-base.info/htb/v8-6-7/Integrase.html> accessed 3-23-2008.
8. Merck Clinical Trials (2006). A multicenter, double-blind, randomized, active-controlled study to evaluate the safety and antiretroviral activity of MK0518 versus Efavirenz in treatment naïve HIV-infected patients, each in combination with TRUVADA™ <http://clinicaltrials.gov/> identifier NCT00369941.
9. National Center for Biotechnology Information Resource Guide, <http://www.ncbi.nlm.nih.gov/Sitemap/ResourceGuide> accessed 3-10-2008.
10. The Stanford University HIV Drug Resistance Database <http://hivdb.stanford.edu/pages/algs/HIVdb.html> accessed 3-23-2008.