

HIVbase[®] 2004

HIVbase Quick Start Guide

The purpose of this exercise is to learn how to use the three major tools in HIVbase. These are the:

- I. Annotation tool
- II. Sequence import tool
- III. Query tool

In each section we provide an overview of the tool and a short practical session to get you started. Importantly, many of the more complicated applications of HIVbase are not mentioned here, however, this basic knowledge will get you started. Use the program for a period of time until you are comfortable with the tools before you compile your own database. Plan in advance what annotations you want to track, how to edit sequences and run queries. HIVbase is a powerful application that is very flexible to your own needs. Above all, enjoy how fast and organized your data processing will become! If at anytime you have questions, do not hesitate to contact us at info@genejohnson.net. We are happy to interact with the research community.

I. Annotation tool. Before you begin to load your own data, you must learn how to set up your own annotations. Annotations are pieces of associated sequence information. Sample ID, date, patient therapy are examples of annotations. Most researchers keep this kind of information in a spreadsheet; however, once you have incorporated this data within HIVbase, you will find that you will be able to quickly put together interesting datasets for further analysis. For example, you may want to look at all V3 envelope sequences from patients that had a CD4 count greater than 500 or determine if a certain sequence motif occurs in all subtype C sequences from patient on a specific therapy. You will be able to run these reports and extract these sequence data sets in FASTA format within seconds.

In this example you will learn how to set up two different annotations. These are "Sample Date," and "Tissues"

Sample Date

1. Open up the program and under "tools" choose "customize annotations."

2. A window will open up called "Customize Annotations Definitions." This window will be empty if you are using a new or trial version of HIVbase.
3. Click the button on the bottom left, "Add..." A window will open up called "Select Annotation Type"
4. Choose the option "Date Values" and click "OK." Now a window opens where you will put information about that annotation.
5. You must give the annotation a name. Type "Sample date" in the name box.
6. Because this is a simple date annotation, you do not need to enter in any more information. Click the, "OK" button and the window will close. You will see that "sample date" now appears in your list of annotations.

Tissues

1. Follow steps 1-3 in the first example.
2. In the "Set Annotation Type" window, choose "text values" and click "OK"
3. In the New Annotation Definition window, type in the name "Tissues."
4. If you would like to provide a detailed description of this annotation you can type whatever you want in the description box. This box is primarily for annotations that later may become confusing, for example, sample date and sequence date.
5. A text annotation could be restricted to certain values or non-restricted. For example Sample ID is unrestricted text because there are so many possibilities, whereas, tissues could be restricted if you know there are a limited number of values. If you want your annotation to be restricted, then check the box on the bottom left of the screen.
6. For Tissues we can add in values. Click on the "Add..." button and type in a tissue name, for example, "Lung." then click "OK" and Lung now appears in the Standard values list. Repeat this step for three more tissues (Blood, Spleen, Thymus).

Examples of other annotations:

1. Subtype (values = A, B, C, AE, D, etc.)
2. Country (values = US, ZM, BW, etc.)
3. Protease Inhibitors (values = Idinovor, ritonavir, etc.)
4. Pathology (values = Kaposis Sarcoma, lymphoma, dementia, none)
5. Living? (values = TRUE, FALSE)
6. Sample ID (No text restriction)
7. Cell Type (values = macrophage, T-cell, ra, ro, etc.)
8. Researcher

II. Sequence Import Tool. HIVbase is designed to extract a lot of information about sequences during import as well as provide you with a means to attach your annotations to each sequence. HIVbase will translate all your sequences in six reading frames and identify HIV proteins and domains. HIVbase will accept ambiguous nucleic acid codes as well as N or X. Importantly, HIVbase will not be able to process "bad" data, such as data

with many frame shifts or many unknown nucleotides or amino acid residues. **YOU MUST STILL PROOF SEQUENCES.** Also, **DO NOT ENTER ALIGNED DATA CONTAINING DASHES.**

In this exercise we will import a sample data set and learn the many functions and kinds of information that can be viewed in the sequence import screens.

Importing data:

1. Download the file "sample_data.fsa" from the HIVBASE website and save it on your computer.
2. In HIVbase, click the File menu on the toolbar, and choose "import." Then, choose to import Fasta Nucleic Acid Sequences." Locate the sample data file and open it. Depending on the speed of your computer and the size of your file, sequence import may take a few seconds or just a little longer. As HIVbase opens your file, it is translating each sequence into six reading frames and identifying proteins and domains of interest (e.g. envelope V3, gag P24, etc.). It is NOT saving the data to your computer. Once the data appears on the screen, you have the ability to add annotations, be sure that the data contains the proteins and/or domains you expect. Be sure that each sequence doesn't have frame shifts, etc.
3. At this point you should see a screen with three windows:

a. SEQUENCE NAME WINDOW. The window on the left contains the name of each sequence imported. If you click on a sequence, it becomes highlighted and it appears in the box at the top of the list. When it is in the top box, you can edit the name of the sequence in the box, if you wish to do so. If a sequence appears in the list that you do not wish to import, highlight it and the click on the box at the top of the list with the red X. The sequence will be removed from the list. Additionally, you will notice that when a sequence is highlighted, the sequence data appears in the window on the bottom right.

b. SEQUENCE VIEW WINDOW. The window on the bottom right is purely an *information screen*. Specifically it shows you much information about each sequence that was calculated during the initial stage of import.

The Sequence Data tab shows the raw data that was imported. Here you can check to see if the sequence looks okay, for example, does it contain a lot of ambiguous amino acids? You can highlight regions of the sequence or the whole sequence to calculate the length (you will see this in the small boxes at the bottom of the window.)

The Nucleic Acids Domains tab provides you with pieces of information on each protein or domain found during import. The first column shows if a particular protein domain occurred in the reverse compliment or in the forward direction. The second column shows which reading frame the domain

occurred in. If, for example, you see domains such as V1 and V3 in different reading frames, you may want to go back and proof that sequence before entering it into the database. The third column shows the major protein the domain was found in. The fourth column shows the domain within that protein that was found. And the fifth column shows the actual sequence of that domain.

The Amino Acid Domains tab is identical to the Nucleic Acids domains window except you are working with the amino acid translation.

In both the Nucleic Acids Domains window and the Amino Acid Domains window you can get further details about a sequence by highlighting the domain (any column in the window will do) and clicking on the details button. The window that opens shows you where the domain is located in the raw sequence and the length of the domain.

c. ANNOTATIONS WINDOW. In the annotations window on the top right you will see all the annotations we created in the first part of the tutorial. You should now see "Tissues" and "Sample Date" and any others you may have created.

Practical exercise:

1. Highlight the sequence A.KE.22.Q842 at the top of the Sequence Name Window.
2. Click on "Tissue" in the annotations window. You will notice that both the sequence name and the annotation are now highlighted. Also, you will see that a drop-down tab is now located at the end of the row that pertains to tissues. Click on the drop down tab and you will see the different tissue values that you have created, like "lung."
3. Choose lung. If you highlight other sequences, and then return to same sequence you have annotated, you will see that "lung" always appears as connected to that particular sequence.
4. Now annotate a sequence for sample date. Because you defined this annotation as a date, when you go to type in a date, you can either hand type a date, or you can use the drop down tab at the right and a calendar will appear.

5. Annotating multiple sequences at the same time. Often the sequences you enter in HIVbase will share values for annotations, for example, if you sequenced 100 clones from the same individual, many values will be the same, including therapy, sequence ID, subtype, origin, treatment, etc. Some annotations will be different, such as sample date, tissue source, etc.

To annotate multiple sequences during an import is quite simple:

- a. if the sequences are grouped together, highlight the top sequence and hold down the shift button while scrolling down the sequence name list.

b. If your sequences are not grouped together, then you can use the control key in combination with a left mouse key to select them. Try to perform this function now with the seven sequences you have on your screen. Select the top sequence, hold down the control key, and click on the last two sequences. When all three are highlighted, go over to the annotation screen and give them a Sample Date annotation.

6. Once you have all the desired sequences with associated annotations assigned to the sequences, **THEN YOU SAVE THEM TO THE DATABASE.** To save all the information, click the "OK" button on the right bottom of the screen. You will get a "confirm import" message in order to verify that you want to save the information. Click OK.

III. The Query Engine. The HIVbase Query Engine is designed to perform three major functions.

1. Retrieve data
2. Edit Data
3. Perform analysis based on amino acid composition

Retrieving data:

1. Go to "File" on the Toolbar and choose "New" and then "Query." A window will open titled "Query Wizard - Step 1 of 3."

In the "Available Fields" column, you will see the annotations you developed and also some that are provided with the program. You may choose as many of these to view in the return set that you want. To get started retrieving sequence data follow these steps:

- a. Highlight "Sequence name" and click "Add." You will see that Sequence name now appears in the "Selected Fields" column.
- b. Highlight "Domain AA data" and click "Add"
- c. Highlight "Domain Name" and click "Add"

At this point you are telling the program to retrieve all sequence names with all of the associated domain data (every protein and domain for every sequence). Click on the "next button at the bottom of the screen.

2. Now you are on a screen called "Query Wizard Step 2 of three." In this screen you can reduce the data set by applying certain conditions. Perform the following steps:

a. In the top "Database Field" box, click on the dropdown tab. Again you will see all of the annotations in the system. Choose "Domain Name" in the list.

b. In the top "operator box," from the drop down list choose "eq." "Eq" stands for "Equals to." The other operators are:

1. contains

2. gt = greater than
3. gte = greater than or equal to
4. lt = less than
5. lte = less than or equal to
6. neq = not equal to
7. is list
8. is null

c. In the value box, click the drop down tab. You will see all options for domain names in the menu. If you had chosen tissues for the database field, all of tissue annotations would be in the values list. Choose V3. Then click "Next."

3. You are now on a screen called "Query Wizard - Step 3 of 3." This step is an information screen and shows you how you have developed the entire query.

Click on "finish" at the bottom of the screen. Now you have opened the query results screen. On the left you have several options, including going back and modifying your query. Choose, "Run Query" from the menu. Now you can see four columns of data. The first column is a select column with small check boxes. Here you can select certain sequences for export, deletion, editing, etc. The second column contains the names of your sequences. The third column contains the name of the domain and the fourth contains the sequence of the v3 domain for each sequence.

Click the "Select All" button on the left and you will see that the options under the "select all" button are now available. At this point you can choose to export all the results into an Excel spreadsheet or into a FASTA file. You could also delete the sequences.

Editing Data

After you have retrieved a data set, you can edit the annotations of many sequences or the sequence itself. Perform the following tasks:

1. Run the query described above and get to your data return set. In the "select column." Rapidly double click in the check box for a specific sequence. This sequence will now be displayed in a new edit screen. You can type in the sequence screen to change the sequence or edit the annotations for that specific sequence.
2. Close the sequence edit screen. Select all sequences. Then click on the edit annotations button on the left. A screen will pop up allowing you to edit the annotations for all of your selected sequences.

Analysis based on amino acid composition

Click on the "Change this Query option." You will return to the first step in the query wizard. There are special functions of the query engine located in the available fields column:

1. Count. This option allows you to count the existence of a particular amino acid or motif in each sequence in your return dataset. For example you can count the number of cysteine residues, glycosylation sites, or the length of a domain.
2. Position. This option allows you to look at a specific position within each sequence in your return data set. For example you may want to look at specific position in protease to see if an amino acid associated with resistance occurs.
3. Exists. This option allows you to find out if a amino acid or motif exists within each sequence in your return data set.

These features allow you to perform very sophisticated queries and will be explained in more detail in an advanced users guide and in a publication under review.

The Query Engine is a powerful tool that allows you to view your data in many different ways. You should allow yourself some time to play around within the different screens and see the many options available to you that have not been explained in detail in this quick start guide.

Please continue to check at www.HIVbase.com for more information as it becomes available, or contact us at info@genejohnson.net with specific user questions.