Search Tools in the HIV Databases

Jennifer Macke, Charles Calef, Karina Yusim, Robert Funkhouser, Thomas Leitner, James Szinger, Brian Gaschen, Werner Abfalterer, John Mokili, Brian Foley, Bette Korber, Carla Kuiken
Los Alamos National Laboratory, Los Alamos, NM 87545
seq-info@lanl.gov http://hiv.lanl.gov

The Los Alamos National Laboratory HIV databases serve as a repository for large amounts of information about HIV. A challenge for us is to find ways to make this information as useful and as easily accessible as possible for our experimentalist colleagues. Over the years, we have developed a variety of web-based tools for searching the databases that we maintain. We hope that by providing a general overview of these search tools, this article will familiarize our website users with new ways of accessing useful data.

Below is a list of our database search programs, followed by detailed descriptions and examples of each.

ALL DATABASES

Site search Find web pages on all our websites.

SEQUENCE DATABASE

Sequence Search Interface Find, align, download, build tree, and analyze sequences.

Advanced Search Interface Generate a custom search interface for any database fields.

Geography Find and map the number of sequences of each genotype by region or country.

IMMUNOLOGY DATABASE

CTL(CD8+) and T-helper(CD4+) Search Find CTL and T-helper epitopes.

Antibody Search Find HIV-specific antibodies.

VACCINE TRIALS DATABASE

Regular Search Search for studies meeting your criteria.

Cross-Table Search Generate cross-tabulated data, based on any two database criteria.

Adjuvants/Stimulants Search a separate database of substances used as adjuvants/stimulants in vaccine trials.

RESISTANCE DATABASE

Simple Search Search for drug resistance mutations by gene, compound, drug class, and amino acid position only.

Advanced Search Search by a wider selection of fields.

ADRA Identify mutations associated with anti-HIV drug resistance in your query sequence.

I. ALL DATABASES

Site Search

The site search can be accessed from the menu bar found on most pages of the Sequence and Immunology databases, shown below. It is a Google-based search for words or phrases in any of the LANL HIV and HCV databases. For example, if you want to know what pages we have about circulating recombinants of HIV, type "CRF" or "recombinants" into the search box. The searches have the same advantages and disadvantages as all Google searches.

II. SEQUENCE DATABASE

The sequence database (www.hiv.lanl.gov/content/hiv-db) contains all HIV-1, HIV-2, SIV, and SHIV sequences that have been deposited in public databases. We have a slight lag-time in retrieving sequences, so some very recent sequence submissions may not be available at the time of search. We provide three search interfaces for extracting sequences and related information: Regular search, Advanced search, and Geography search.

Regular Search Interface

The information in the sequence database can be accessed via a versatile, user-friendly search interface that allows searches on approximately 30 different fields. An important feature is the ability to search by genomic region. For example, you can locate all sequences in the database that span the V1–V3 region of *env*. There is also an option to include sequences that are located in that region but do not cover it completely; this option is labeled "Include fragments of minimum length __". The minimum length specified is the length by which a sequence needs to overlap the region of interest in order to be included. In the example shown below, we are using the regular search interface to select sequences of subtype C from the US. We have further limited the search to only sequences from patients with >2 sequences available in the database, and to the Pol region, including fragments over 100 bp long.

Fields available on the Search Interface Mousing over the names of most fields gives a very brief description of each field; more help is available by clicking on the field name. Seven fields are always included in the output list: accession number, sequence name, subtype, sampling country, sampling year, genomic region, sequence length, and organism. All additional fields that are included in the search are also listed in the output. (Note that this can result in wide pages for some searches, such as Author names.) A complete description of all database fields is found in the search interface help file (www.hiv.lanl.gov/components/hiv-db/combined_search_s_tree/help.html). We give brief descriptions below.

Accession number To search for a range of accession numbers, type X12345 .. X23456. You cal also search on a part of the number: X1234 gives you all accession numbers that start with this string.

Subtype PC users can select multiple subtypes by using 'ctrl-click.' For most other browser/platform combinations, either shift or command will do this. If you are interested in subtypes not included in this list, use the Advanced Search interface, described below.

Sequence Information

Accession number

Sequence name

Sampling country US

Sampling city

Sampling year

Sequence length

Subtype
no subtype
A
A1
A2
B
C

incl. recombinants ☐

Organism Any

incl. problematic seqs ○ yes ● no

Patient Information

Patient code

Risk factor Any

Viral load

CD4 count

Infection country

Infection year

incl. only sequences from patients with HLA information ○ yes ● no

CD8 count

Other Information

Coreceptor
list field in output
only CCR5
all CCR5
only CXCR4
all CXCR4
only R5X4

Author last name

Geographic region Any

Pubmed/Medline ID

Phenotype
list field in output
NSI
SI

incl. only drug naïve sequences ○ yes ● no

Days from seroconversion

Days from infection

# of patient seqs >2

Find all sequences for a specific gene or region

Specific regions of HIV-1 and SIVcpz can be retrieved from the HIV database using the following options:

☑ Include fragments of minimum length 100

Genomic region
5' LTR U3
5' LTR U5
TAR
Gag–Pol
Gag
p17 (matrix)
p24 (capsid)
p7 (nucleocapsid)
p6
Pol CDS
p51 (RT)
p15 (RNAse H)
p31 (integrase)
protease

Or define start _____ and end _____

To retrieve HIV-1 sequences form the database and integrate them into your own sequence alignment, paste your aligned nucleotide sequences into the submission box below.

Or upload your alignment file:

Browse...

Output

List 100 records per page

Search   Reset

Include recombinants By default, recombinants containing fragments of the selected subtype(s) are included in the retrieval; uncheck this box to exclude them.

Authors Searches for author(s) listed on the publication. Do not include initials. You can also search on part of the name, *e.g*., "James" will find authors with last name James or Jameson.

Pubmed/Medline ID The search returns all sequences associated with a specific PubMed ID. Although Medline IDs are no longer being assigned, they may still be searched and retrieved using this same field.

Patient code The patient code is displayed as a 2-part number, for example "P1(10139520)". The first part is usually the name or number by which the patient is identified in publication(s). The second part is a unique number assigned by our database, the patient ID. A patient code such as "P1" can (and does) refer to more than 1 patient. However, the sequence records associated with "10139520" are

specific to a unique patient.

Risk factor This field describes the risk activity by which the patient most likely was infected.

Infection country We use the official two-letter country code for the infection country. A list of these codes is available on our website.

Infection year The year in which the patient was infected.

Days from infection The number of days from the time the patient was infected until the sample was taken for sequencing.

Days from seroconversion The number of days between the patient's seroconversion and the date the sample was taken for sequencing.

Sampling country The 2-letter code for the country in which the sample was taken.

Sampling year The year in which the sequenced sample was taken.

Geographic region This is a way to retrieve all sequences from (for example) the African continent without having to search for each country separately. Clicking on the field name shows a list of which countries are included in each region.

Genomic region search The user can specify which genomic region to include in the search in three different ways: 1) by predefined region in a table; 2) by HXB2 coordinates; 3) by automatic matching start and stop coordinates to a user provided alignment. Part or all of the sequences are used if they fall within the selected region. A genomic map showing the regions is available on the website. The external sequences are automatically aligned to the search.

Exclude problematic sequences This option excludes certain sequences from results:
- N: high non-ACTG content (N's or IUPAC codes)
- C: potential contamination, as determined by the database staff
- H: hypermutated
- S: synthetic sequence

When searching for sequences by accession number, problematic sequences are included in the results, but their problematic code is indicated; we assume you specifically want these sequences. For other searches, they are excluded by default, but you can include them by unchecking the corresponding boxes.

Other fields This is a pull-down menu to access several other search fields in the database without cluttering the interface. Only one option can be chose per search.

> *Infection city* The city (or region) where the patient was infected.
> *Title* Words from the title of the publication associated with the sequence.
> *Comment* Words in the comments entered by database staff.
> *Patient sex* M or F.
> *Patient age* The patient's age in days at the time of sampling.
> *Project* The name of the project or cohort.
> *Progression* EC, LTNP, SP, RP, or P.
> *Number of patient sequences* Use this field to find patients who have more than # sequences.
> *Patient health* Acute infection, asymptomatic, symptomatic, AIDS, or deceased.
> *Isolate name* Isolate name as given by the authors.
> *Clone name* Clone name as given by the authors.
> *Sample tissue* Material from which the virus was isolated.
> *Culture method* Uncultured, primary [culture], or expanded [culture]

Search results The search results are presented in a table showing some basic information about each sequence. A small graphic for each sequence shows where in the genome it is located. This can be very useful to determine which region is best represented in that set, and therefore most suitable for further analysis.

The search results can be sorted and selected in various ways. Retrieved sequences can be downloaded as an aligned file in FastA or other formats. These alignments need manual inspection and often improvement, but form a very useful starting point. Alternatively, sequences can be downloaded as unaligned nucleotides and/or translated to amino acids in any reading frame.

| Make Tree | Download Sequences | Save Background Info | Clear |

Displaying 1 - 32 of 32 sequences found

Select all | Unselect all | Invert selection | One sequence/patient | Select record [____] to [____] List 32 records per page

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession	Name	Subtype	Country	Sampling Year	# of patient seqs	Genomic Region	Sequence Length	Organism
1	☐ Blast	610185(10140292)	AY140820	610185H	C	US		8		1188	HIV-1
2	☐ Blast	610185(10140292)	AY140821	610185H	C	US		8		1180	HIV-1
3	☐ Blast	610185(10140292)	AY140822	610185H	C	US		8		1180	HIV-1
4	☐ Blast	610185(10140292)	AY140823	610185H	C	US		8		1189	HIV-1
5	☐ Blast	27_5016(10148221)	AY444801	98US_MSC5016	C	US	1998	2		8415	HIV-1
6	☐ Blast	T2(10149045)	AY484347	01UST21tca	C	US	2001	25		559	HIV-1
7	☐ Blast	T2(10149045)	AY484348	01UST21tcb	C	US	2001	25		559	HIV-1
8	☐ Blast	T2(10149045)	AY484349	01UST21tcc	C	US	2001	25		559	HIV-1
9	☐ Blast	T2(10149045)	AY484350	01UST21tcd	C	US	2001	25		559	HIV-1
10	☐ Blast	T2(10149045)	AY484351	01UST21tce	C	US	2001	25		559	HIV-1

The search interface also allows you to download the data in the output table as a tab-delimited file, optionally including the unaligned or aligned sequences. The data can easily be imported into a text editor or a spreadsheet such as Excel.

Sorting the sequences To sort the sequences on the content of one of the columns, click on the title of that column. Clicking again will sort them in the reverse order.

Selecting sequences You can select sequences by checking the boxes at the beginning of the line. To simplify the process, you can also use the blue-on-white buttons at the top of the table. Even if your results are not displayed on a single page, these buttons work across pages. The 'Select all' and 'Unselect all' functions are obvious. Use 'Invert selection' when you want to exclude a few sequences; select those and then invert the selection. That will save a lot of clicking. 'Show all' allows all sequences to be listed on one page. You can use 'Select record __ to __' to select a range of sequences; the numbers refer to the line numbers in the table. Finally, use 'List __ records per page' to change the length of each page.

Limiting the set to 1 sequence per patient This button is only displayed when a genomic region is selected as one of the search criteria. It randomly selects one sequence from all sequences in the search result that share a patient record. In other words, if there are multiple sequences that are known to be from the same patient, all but one are discarded. Note that if multiple sequences per patient are present but no patient record exists, these sequences will be deleted from the set. In other words, this function is dependent on our annotation.

Downloading the sequences aligned vs. unaligned If your set contains only HIV-1 sequences, you can download nucleotide sequences as an alignment, or unaligned. HIV-2 and SIV sequences cannot be pre-aligned because no pre-aligned sequences are stored in the database. Amino acids only come un-aligned, in any (or all) of the 3 reading frames.

If you used a genomic region or sequence coordinates to retrieve your alignment and you have checked the 'clip to selected region' box, your sequences will be limited to the selected region. Other-wise, you will end up with an alignment that covers the entire genome including the alignment gaps, i.e. is around 11,000 characters long. This can be convenient if you want to align your sequences to a set of complete genomes, or to other sequences retrieved using the same method (these alignments may differ by a few positions). Note that the alignments are not necessarily optimal and usually require manual adjustment. If you download an alignment, sequences that do not have valid coordinates relative to the reference sequence will not be included in the alignment. Aside from HIV-2 and SIV sequences, this can also happen if the sequences are very short, if they contain non-HIV inserts, or if they are reverse complements. These sequences will be easily noticed in the search interface output because they do not have the icon that shows the location, but say "no location info" instead.

Including a reference sequence or reference alignment You can include the reference sequence HXB2 in your downloaded sequences. This will make it easier to use the SynchAlign tool to align these sequences to other sets.

How the sequences are aligned When the sequences are uploaded into the database, they are internally aligned against a 'model sequence' that represents all sequences that are already present in the database. For this alignment, we use the HMMER program, written by Sean Eddy (http://hmmer.janelia.org/). The start and end coordinates of each sequence relative to the model sequence, as well as the location of all the gaps, are stored in the database. When you request all sequences encompassing the *vif* gene, for example, the coordinates for the *vif* gene in the model sequence are retrieved, and all sequences with a lower (or equal) starting point and a higher (or equal) stopping point are retrieved. When the sequences are downloaded, the gaps relative to the model sequences are inserted. For the little image that shows the location of the sequence relative to the genome, a slightly different set of coordinates is used, relative to the reference sequence (HXB2 or SIVsmm239) instead of the model sequence. These coordinates are produced by an algorithm, and are identical to the coordinates that the Sequence Locator tool produces. The location of some sequences cannot be accurately determined, often because they are too short. In these cases, the sequence will not be included in the aligned download, but if you download the sequences unaligned it will be there.

Creating a phylogenetic tree You can make a neighbor-joining tree from all or a subset of your retrieved sequences, your aligned user sequences, and include subtype reference sequences. The interface allows you to compose labels for your sequences, to choose the evolutionary model for the distance calculation (currently F84, Jukes-Cantor, Tamura/Nei, Kimura 2-parameter, and the General Reversible model), to set gap handling options, to set site rate variation, and to choose the outgroup sequence. The alignment, treefile, and various graphical representations of the tree can be downloaded.

Downloading background information It is possible to download the output shown in the search results table as a tab-delimited file, which allows you to tabulate background data for the retrieved set. Examples of background information: patient information (code, health status, age, gender, risk factor, infection date, infection country, viral load), comments (from the authors or the database staff), and sequence information (sampling city, clone name, etc). These fields can be from any of the fields available from the search interface.

Links within the search results table Several links are located inside the search results table. These links only apply to the sequence on that line.

BLAST does a search of the sequence against the HIV database.

Accession Clicking on this link displays a "GenBank-style" entry that contains all data from GenBank, plus extra features added by our database. You can download the entire sequence or part of it in several formats; there is a link to the original NCBI entry; and there are links to "Show all sequences for reference X". These links use the publication ID to retrieve all sequences that are associated with that publication or sequence deposit. Note: if you wish to display or save the GenBank information for your whole set of selected sequences, go to the "Download sequences" option and choose "GenBank" as the format.

Patient Clicking the patient ID displays all the information available in the databasefor that patient with links to sequences drawn from that patient.

Genomic Region Mousing over the little green-and-yellow icon shows the exact start and stop coordinates of the sequence.

Advanced Search Interface

This interface dynamically reads the schema of the database and generates a graphical overview of the tables and fields. You can use this overview to generate your own custom-made search interface. Just check the boxes next to the fields you want to either search on, or list in the output. However, some of the 'overhead' that the regular interface performs automatically must be done by hand in the advanced interface. For example, when you check fields from multiple tables, you need to make sure that they share a key (shown in red), otherwise the search will fail. You can get information and examples of the content of the tables and fields by mousing over or clicking on the table names.

Make [Search Interface]   [Reset]

Accession(SA) ☐
☑ SE id(SA)
☑ Accession

Sequence sample(SSAM) ☐
☑ SE id(SSAM)
☑ PAT id(SSAM)
☑ Name
☐ Locus name
☐ Isolate name
☐ Clone name
☑ Country
☐ Sampling city
☑ Sampling year
☐ Patient Age
☐ Patient health
☐ Organism
☑ Subtype
☐ Phenotype
☐ Coreceptor
☐ Sample tissue
☐ Culture method
☐ Sequencing method
☐ Molecule type
☐ Georegion
☑ Drug naive
☑ Problematic
☐ Viral load
☐ CD4 count
☐ CD8 count
☐ Sampling year delta
☐ Days post infection
☐ Days from seroconversion

Patient(PAT) ☐
☐ PAT id
☐ Patient code
☐ Patient sex
☐ Risk factor
☐ Infection country
☐ Infection city
☐ Infection year
☐ Patient comment
☐ HLA type
☐ Project
☐ Patient ethnicity
☐ Infection year delta
☐ Progression
☐ Patient cohort
☐ # of patient seqs

In the example shown here, we are going to make a search interface where we can retrieve all sequences covering Pol protease and RT (coordinates 2253–3870) from patients that are known to be non-drug-naïve. We have checked several additional fields so that we can view these fields in the output.

When you have made your selection of fields, click the "Search interface" button, and a search interface will be generated, shown on the following page. This interface looks much like the standard search interface, but it will contain exactly the fields you selected. You can search on any of the fields, and all fields included in the search interface will be listed in the output. You can also set the number of search results you want to list per page.

The results page from the Advance Search looks similar to the results page from the Regular Search, but with fewer options. You can sort on any of the fields, and download the sequences and/or the background information. Note in the example shown, that including the "Problematic" field in the search allows us to see that one of the sequences is a hypermutant.

Custom search interface (above) based on the selections specified on the previous page. The results page (below) is similar to the standard results.

Advantages of advanced search interface:
• Advanced search allows additional types of searches not possible on the regular interface. As one example, you could restrict your search to include only samples from non-drug-naive patients.
• You can search and view data from all available fields, including multiple fields that appear under "Other Fields" in the regular interface.
• You can specifically select sequences that have a "null" value (no data) in specific fields.

Limitations of advanced search interface:
• The search output does not directly interface with TreeBuilder. To make a phylogenetic tree, export your aligned sequences in FastA format, then use them as input in the TreeMaker tool.
• Options for searching for a specific genomic region are limited.
• Problematic sequences are not removed by default. You will not see any indication of which sequences are problematic unless you include this field in your search.

Other differences from the standard search interface:
• Searches are limited to 10,000 results. If your search produces more than this, it will fail. Restrict your search to produce fewer results.
• To retrieve sequences from part of the genome, *e.g.*, all *vpu* sequences, you need to find the HXB2 coordinates for *vpu* (6062–6310), and then search for HXB2 start < 6062 and HXB2 end> 6310. It is not possible to include sequence fragments smaller than that range.
• The advanced search interface is case-sensitive; it does distinguish between lower- and upper-case letters. When searching on text fields, if you unexpectedly get no hits, try UPPERCASing and/or adding an * to the search. Adding an * will turn the search into a case-insensitive wildcard search. Note that the * will only expand in its own location, so if you want to search for a string in the middle of two unknown other stings, use *string*.
• When you generate your customized search interface, all fields are pre-filled with "ANY". This will display what is entered in that field (even if blank), but will not restrict the search. If you remove the word "ANY", you will restrict the search to entries that have no data in the field (see next point).

Geography Search Interface

The Geography tool is another way to select sequences from the database. It can be used to find the number of sequences of each genotype within any selected geographical region. The information can be extracted as a graphic map, as a table of data, or as a list of specific sequences. The list of sequences connects with the search interface to allow rapid retrieval and analysis of the sequences that are displayed. This tool can be very useful to get a general idea of what genotypes have been found in what countries, as well as the density of sampling in different regions of the world.

Data can be extracted for the whole world, for a region, or for any specific country. The regions available to search are: Africa, Asia, Caribbean, Central America, Europe, Former USSR, Middle East, North America, Oceania, South America, and Sub-Saharan Africa. Note that some of these regions are overlapping, such as Africa and sub-Saharan Africa. A complete list of what countries are included within each region is available (hiv.lanl.gov/content/hiv-db/HelpDocs/geo_regions.html).

Searches can be limited to either HIV-1 or HIV-2. Data can be extracted for the whole world, for a region, or for any specific country. Search results include all sequences, regardless of their length or location within the HIV genome. Search results exclude sequences for which the database lacks an annotation of subtype or country, and sequences annotated as "problematic".

Performing a search To run the search, select HIV-1 or HIV-2 and the desired geographic region. Only 1 region or country may be selected for each search. Click "Show all", "show non-recombinant" or "show recombinant". The resulting page displays a pie chart of the sequence subtypes in the region or country of interest. From here you have the following options:
• View the pie chart. You can use the "save image" function of your browser to save this image as a PNG file, if desired.
• Click on an individual country within a region to obtain the data from that country (this applies only to searches of regions).

• Click on "Table (html)" to go to a page with a table showing the number of sequences with each subtype.

• Click on "Table (text)" to download a space-delimited text file of the subtype distribution data. This text file can be opened in a spreadsheet program.

• Click on "get all" or click on any specific slice of the pie chart. These options take you to a list of the sequence accession numbers. This list is in exactly the same as the list of sequences you would get from the regular sequence Search Interface. From this list, you have many options: Make Tree, Download Sequences, or Save Background Information. The options from this page are explained above in detail in the explanation for the Regular Search Interface.

Examples of geography searches Perhaps the best way to understand the possibilities offered by this interface is to look at some specific examples of hypothetical questions you could answer with this tool.

1. *What are some CRFs that commonly occur in sequences from Brazil?* Select HIV-1/Brazil and click the "Show recombinant sequences" button. The resulting pie chart shows just the CRF and other recombinant sequences sampled in Brazil. To see what subtypes are represented in the "Other" category, click the corresponding pie slice, and you will see them listed.
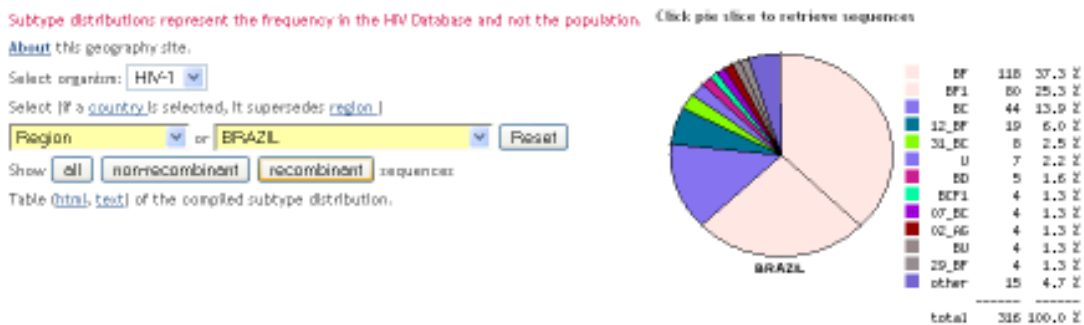
2. *What are the rarest HIV-1 sequence subtypes ever sampled in Sweden?* This is a search that would be difficult using the regular Search Interface, but easy here. Select HIV-1/Sweden and "Show all". In the resulting pie chart, click on the "Other" slice. The resulting list (see below) shows the subtypes of the sequences that were too rare to display as separate pie slices.

Limitations and additional details There is some redundancy in the information that can be extracted by the Geography Search and the regular Search Interface. For example, if you want to extract all subtype B HIV-1 sequences from Africa, you could use either interface. The main reason that you may prefer to do the search using the Geography Search is to see a graphic presentation of all the genotypes. The graphic output may provide some interesting insights that you would not notice in the lists obtained from the sequence Search Interface. However, if the objective is only to extract sequences of a single subtype, regardless of the geographic distribution of that subtype or the representation of that subtype relative to others, then the regular Search Interface may be the better option.

The results from this tool need to be interpreted with care: it is easy to overlook the sampling biases that can distort the frequencies of sequences in the database relative to those in the population. Do not draw conclusions about the epidemiology of HIV-1 from the subtype distribution presented here. The data stored in the database are taken from publications in the literature, and in general, there is no epidemiological framework - the database is just a listing of available sequences. Many studies focus on rarer subtypes and recombinants, and this tends to cause overrepresentation of such sequences. Furthermore, the distribution shown on the maps is based only on the country of sequence isolation, which is not always the country of infection. While the results of this tool are not particularly helpful for epidemiological purposes, one can still use the tool to get a sense of how intensively a region has been studied and a rough indication of the subtype distributions.

The data generated by this search are only as good as the annotation of the sequences in the database. There can be errors in the subtyping of sequences, so all results should be interpreted carefully. When examining rare subtypes, it may be worthwhile to verify the correct subtyping of specific sequences. Furthermore, not all the sequences in the HIV sequence database have an assigned subtype. At the time of this writing, approximately 15% of all sequences in the database have no annotation of subtype. Sequences where the subtype field is blank are not reported in the output from this tool.

Small sample sizes can also be problem with this tool. In countries where there are few subtyped sequences, one or a few studies (with whatever their objectives may have been) will determine the output. This is particularly true for HIV-2, for which far fewer sequences are available than for HIV-1.

This tool lumps together sequences that are annotated with certain sub-subtypes. For example, the sequences listed as subtype "A" from this tool include all sequences annotated as "A", "A1", or "A2". Sequences listed as subtype "F" from this tool include all sequences annotated as "F", "F1", or "F2". However, sub-subtypes that are part of recombinants are not lumped together. For example, "BF", "BF1", and "BF2" are graphed separately.

III. IMMUNOLOGY DATABASE

The HIV immunology database provides resources for scientists working with immunological responses to HIV. The database contains a wealth of curated information about HIV T-cell epitopes and antibody binding sites. Currently the HIV immunology database contains 3818 cytotoxic T-cell (CTL) epitope entries, 829 T-helper epitope entries, 1366 antibody entries, and a total of 1895 publications. New entries are added and proofread continually, and existing ones are updated as needed.

The data included in HIV immunology database are extracted from published HIV immunology literature. HIV-specific B-cell and T-cell responses are summarized and annotated. Immunological responses are divided into 3 sections of the database: CTL (CD8+), T helper (CD4+), and antibody (Ab). Within these sections, defined epitopes are organized by protein and binding sites within each protein, moving from left to right through the coding regions of the HIV genome. We include human responses to natural HIV infections, as well as vaccine studies in humans and a range of animal models. Responses that are not specifically defined, such as responses to whole proteins or monoclonal antibody responses to discontinuous epitopes, are summarized at the end of each protein sub-section. Studies describing general human responses to HIV, but not to any specific protein, are included at the end of each section.

The annotation of database entries includes information such as cross-reactivity, escape mutations, antibody sequence, TCR usage, functional domains that overlap with an epitope, immune response associations with rates of progression and therapy, and how specific epitopes were experimentally defined. Basic information such as HLA specificities for T-cell epitopes, isotypes of monoclonal antibodies, and epitope sequences are included whenever possible. In addition, the HIV immunology database includes tables, maps, alignments of HIV-specific CTL, helper and antibody epitopes, antibody indices, and simple web-based tools with the goal of assisting immunologists in experimental design and interpretation of their results.

An important distinction between the T-cell and antibody entries is that a single T-cell epitope can have multiple entries; generally each entry represents a single publication. In contrast, each monoclonal

antibody (MAb) has a single entry that includes all publications we could find that refer to the use of this specific monoclonal antibody.

CTL (CD8+) and T-helper (CD4+) Searches

CTL (CD8+) and T-helper (CD4+) epitope database sections are organized identically and so are described together here. It is important to note that although these are separate sections, the simple distinctions between CTL and helper T cells have become blurred as more is learned about the range of responses triggered in CD4 and CD8 positive T-cells. When adding the most recent studies, we have tried to place T-cell responses in a reasonable manner into our traditional T-helper and CTL sections, and to specify the assay used to measure the response in each study.

The following is a list of fields and links in T-cell epitope entries:

Record number A unique number assigned by the database, in approximate order of entry. This number should be cited if you send us comments or questions about an entry.

HXB2 Location The position of the defined epitope location is given relative to the protein sequence of HXB2. Because of HIV-1 variation, the epitope may not actually be present in HXB2, rather the position in HXB2 indicates the position aligned to the epitope. The viral strain HXB2 (GenBank Accession Number K03455) is used as a reference strain throughout the database. HXB2 was selected as the reference strain because so many studies use HXB2, and because crystal structures for HXB2 proteins are available. The precise positions of any epitope relative to the HXB2 reference strain can be readily obtained using the interactive position locator at our web site: http://hiv-web.lanl.gov/content/hiv-db/LOCATE/locate.html

Author Location The amino acid positions of the epitope boundaries relative to the reference sequence are listed, as given in the primary publication. Frequently, these positions are imprecise or are based on a non-HXB2 strain. Thus, these locations do not always match the HXB2 numbering of the sequence, but they provide a reasonable guide to the peptide's approximate location in the protein.

Subtype The subtype under study; generally not specified for B subtype.

Epitope Sequence The amino acid sequence of the epitope of interest as defined in the reference, based on the reference strain used in the study defining the epitope. On occasions when only the position numbers and not the actual peptide sequence was specified in the original publication, we try to fill in the peptide sequence based on the position numbers and reference strain. If the sequences were numbered inaccurately by the primary authors, or if we made a mistake in this process, we may misrepresent the epitope sequence. Because of this uncertainty, epitopes that were not explicitly written in the primary publication are followed by a question mark in the table.

Epitope Name The epitope's name if attributed by the publication, *e.g.*, "SL9".

Species (MHC/HLA) The species responding (*e.g.*, chimpanzee, mouse), and MHC or HLA specificity of the epitope, as described in the primary publication (*e.g.*, A*0201).

Immunogen The antigenic stimulus that generated the initial immune response, *e.g.*, HIV infection or vaccine. If a vaccine stimulated the response, additional fields are available that describe the vaccine.

Vaccine type The vaccine construct and boost.

Vaccine strain The strain of HIV or SHIV used for the antigen.

Vaccine component The HIV protein (complete or partial) included in the vaccine.

Adjuvant Stimulatory agent sometimes included in a vaccine formulation to enhance or modify the immune-stimulating properties of a vaccine.

Country The country from which the samples were obtained; generally not specified if the study was conducted in the United States.

Experimental methods The methods used by the authors to test the immune response (for example T-cell Elispot, intracellular cytokine staining, etc.).

Keywords A searchable field for the web interface to help identify entries of particular interest.

Notes Brief descriptions of what was learned about the T-cell response from the study. Examples of the kinds of things included: correlation with survival in longitudinal studies, immune escape, quantitative features of the response, subtype cross-reactivity, etc.

Reference The primary reference linked to PubMed.

The CTL and T-helper search interfaces, shown below, allow you to search using the following fields: HIV protein (separately for defined and undefined epitopes), epitope sequence, subtype, immunogen, vaccine details, species, MHC/HLA, author (any one author from primary publication), country, and keywords.

HIV Immunology CTL, CD8+ T-Cell, Search

HIV Protein	Proteins with defined epitopes	Proteins with undefined epitopes		
	– ALL – p17 p17–p24 p24 p24–p2p7p1p6	– ALL – Gag Gag/Pol Protease–RT Pol		
Epitope				
Subtype	– ALL –			
Immunogen	– ALL – computer prediction HIV–1 and HCV co–infection HIV–1 exposed seronegative HIV–1 infected monocyte–derived HIV–1 infection HIV–1 or HIV–2 infection			
Vaccine details	If Immunogen is Vaccine, additional search details			
	Vaccine type	– ALL –		
	Vaccine strain	– ALL –		
	Vaccine component	– ALL –		
	Adjuvant	– ALL –		
Species	– ALL –			
MHC/HLA	– ALL – – NULL – A*0101 A*02 A*0201 A*0201, A*0205 A*0201, B*3501			
Author		☐ First ☐ Last		
Country	– ALL –			
Keywords	– ALL – acute/early infection adjuvant comparison antibody generation assay standardization/improvement binding affinity characterizing CD8+ T cells			

Antibody Search

The antibody database summarizes HIV-specific antibodies (Abs) arranged sequentially according to the location of their binding domain. Monoclonal antibodies (Mabs) that do not bind to defined linear peptides are grouped by category at the end of each protein. Antibody categories (e.g., CD4 binding site (CD4BS) antibodies) are also noted in an index in the compendium. Studies of polyclonal Ab responses are also included. Responses that are just characterized by binding to a protein, with no known specific binding site, are listed at the end of each protein section in the compendium.

Each MAb or polyclonal response has a multipart basic entry that includes all publications that refer to the use of that specific Ab. Most of the fields are similar to the corresponding fields for T-cell epitopes, such as record number, HXB2 location of the binding site, author location, epitope sequence, and immunogen (the antigenic stimulus of the original B-cell response). The fields that differ or are

specific to antibody entries include the following.

MAb ID The name of the monoclonal antibody with synonyms in parentheses. MAbs often have several names. For example, punctuation can be lost and names are often shortened (M-70 in one paper can be M70 in another). Polyclonal responses are listed as "polyclonal" in this field.

Neutralizing L: neutralizes lab strains. P: neutralizes at least some primary isolates. no: does not neutralize. No information in this field means that neutralization was either not discussed or unresolved in the primary publications referring to the MAb.

Species(Isotype) The host that the antibody was generated in, and the isotype of the antibody.

Donor Information about who produced the Ab, where to obtain it, and to whom to provide credit.

References All publications that we could find that refer to the use of a specific Ab.

Notes Describe the context of each study and what was learned about the antibody.

The antibody search interface, shown below, allows you to search using the following fields: HIV protein (separately for defined and undefined epitopes), epitope sequence, subtype, immunogen, vaccine details, species, MAb ID, author (any one author from primary publication), Ab type, country, and keywords.

HIV Immunology Antibody Search

IV. VACCINE TRIALS DATABASE

The HIV/SIV Vaccine Trials Database (www.hiv.lanl.gov/content/vaccine/home.html) contains data on studies of HIV/SIV vaccine trials in nonhuman primates. This database is a tool for compilation, search, and comparison of published studies. We use a set of criteria to scan PubMed for relevant studies to enter into the database. In selecting studies for entry, priority is given to recently published studies in journals generally regarded as the primary source of information pertaining to HIV and SIV vaccine research in nonhuman primates. In most cases, we give priority to challenge studies, where the animals received a live virus to measure the 'efficacy' of the immunogen(s) inoculated during the course of the investigation.

Prior to the development of this database, Dr. Jon Warren at the EMMES Corporation had maintained a similar database, though organized differently, and with different data fields and somewhat different nomenclature. The studies in that database include many published through 1999 and can be accessed in the current Los Alamos Vaccine Trials Database. All of the search criteria on the search form apply to the studies entered at Los Alamos (referred to as the "Current Database"), and *most* criteria apply to the data collected by Jon Warren (referred to as the "Previous Database").

The vaccine trials database can be searched in 2 ways, via a conventional search form, and via a cross-table form. In addition, there is a separate search tool to access a database of vaccine adjuvants.

Search Form

The conventional search form allows you to select trials according to the combined values of 14 separate criteria such as vaccine type, vaccine route, challenge strain, etc. The matching trials can then be displayed in various formats, and detailed information about the trial, such as details on the substances and groups used, are also available. To display all available results, be sure to click "View Trials in Previous Database".

Cross-Table Form

The cross-table form provides a unique way to explore trial data by presenting counts and results (number of case animals protected from infection and total number of case animals) by the relationship of two selected values displayed in a matrix, for example vaccine type by year.

The results will show how many trials were done for each vaccine type in each year. Clicking on the number in the matrix takes you to a list of those trials that compose a result cell.

Vaccine Type x Year of Publication
5/25/07 9:37

Vaccine Type	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Cell/Tissue	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2 [4/21]
DNA	0	0	1	0	2 [5/6]	0	0	0	0	0	2 [3/2]	0	4	12 [8/57]	13 [15/28]	13 [4/66]	12 [0/38]
Live Attenuated Virus	0	0	0	1 [0/3]	0	1 [4/6]	0	0	0	2 [3/12]	2 [1/4]	1	7 [23/60]	11 [1/28]	3 [1/6]	2 [7/12]	5 [4/27]
Live Virus	0	0	0	0	0	0	0	0	0	2 [7/8]	1	0	1	2 [4/8]	3 [7/18]	1 [0/8]	2 [0/28]
Other	0	0	0	0	1 [2/2]	0	0	0	0	0	0	0	0	1	0	2 [3/16]	0
Passive Antibody	1	1 [0/4]	0	0	2 [0/11]	2 [2/2]	0	0	0	0	1 [3/4]	0	6	3 [8/14]	2 [0/18]	2 [10/15]	3 [6/27]
Purified Viral Products	0	0	1	0	0	2	0	0	0	0	2	0	4	3 [1/27]	3 [6/21]	4 [1/26]	3 [0/18]
Recombinant Live Attenuated Virus	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Recombinant Subunit Protein	0	1	0	1	2 [1/5]	1	1	0	1 [1/2]	4 [8/17]	0	4 [5/11]	2 [13/24]	4	5 [8/28]	2	5 [0/28]
Recombinant Vector (virus/bacteria)	0	0	0	0	1	2	0	0	0	1	2	2 [1/3]	6 [15/28]	9 [5/60]	10 [2/34]	14 [2/68]	17 [0/83]
Synthetic Protein/Peptide	0	0	0	0	0	0	0	0	0	0	1	1 [4/8]	0	2 [0/2]	4 [2/28]	1 [0/8]	1
Virus-like Particle	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1 [3/10]	1
Whole (killed) Inactivated Virus	0	0	4 [13/22]	4 [4/16]	0	3 [3/10]	3 [3/11]	0	0	0	0	2	2	0	0	0	1

Like the search form, this search directly retrieves only sequences in the "Current" database; the number presented in each cell of the table represents only the studies in the current database. In order to view entries from both the current and previous databases, click the number in the cell. Even if the number in the cell is 0, you may still find some relevant studies by clicking the cell and going to the "View Trials in Previous Database" link.

Adjuvant/Stimulant Search

The vaccine database also contains a database of adjuvants used in vaccine trials. To access this information, select the "Adjuvants/Stimulants" link from the main menu. Links from the resulting table take you to details about the structure, properties, uses, and references for each substance. The example below shows some of the information given for the adjuvant alum.

Adjuvant / Stimulant: Alum

Other Names: Alhydrogel; Aluminum hydroxide gel;

Structure: Crystalline aluminum oxyhydroxide AlOOH, known mineralogically as boehmite. The structure consists of corrugated sheets of aluminum octahedra.

Adjuvant Properties: Alhydrogel is the standard preparations for immunological research on aluminum hydroxide gels. The use of aluminum adjuvants is accompanied by stimulation of IL-4 and stimulation of the T-helper-2 subsets in mice, with enhanced IgG1 and IgE production. Further immunomodul ation is accomplished by the aluminum content. Properties are described in:
• Shirodkar. S., et al., 1990, Aluminum compounds used as adjuvant in vaccines, Phar7n. Res. 7: 1282 1288.
• Stewart-Tull, D. E. S., 1989, Recommendations for the assessment of adjuvants (immunomodulators), in: Immunological Adjuvants and Vaccines (Gregoriadis, G., Allison, A. C., and Poste. G., eds.), PlenumPress, New York, pp. 213-226.
• Gupta, R., et al.. Chapter 8, this volume. 9
• Seeber, S., et al., 199 1, Predicting the adsorption of proteins by aluminum-containing adjuvants, Vaccine 9: 201-203.
• Seeber, S. J., et al., 199 1, Solubilization of aluminum-containing adjuvants by constituents of interstitial fluid. J. Parenteral Sri, Tech. 45: 156-159.
• Hern, S., and White, J. L., Chapter 9, this volume.
• Lindblad, E. B. Aluminum adjuvants, in: The Theory and Practical Application of Adjuvants, chapter 2, (Stewart-Tull, ed.), Wiley & Sons, New York, NY (1995).

Source: Obtained by precipitation of aluminum hydroxide under alkaline conditions.

Uses: Human applications: diphtheria, tetanus, and pertussis vaccines. Veterinary vaccine applications.

Appearance: White gelatinous precipitate in aqueous suspension.

Molecular Weight: Not applicable.

Recommended Storage: 4-25°C. Never expose to freezing. Recommended 2 year shelf life.

Chemical/Physical Properties: Primary particles have a rodlike or fibril morphology and a high surface area. The isoelectric point is 11. Its high surface area gives it a high adsorptive capacity for antigen. Poorly soluble in solutions containing citrate ions. Normal particle size range 0.5-1 0pm.

Incompatibility: Dissolves in strong bases and acids.

Safety/Toxicity: May cause mild local reactions at the site of injection (erythemas and/ or mild transient swellings).

V. HIV DRUG RESISTANCE DATABASE

The Los Alamos HIV Drug Resistance Database (http://resdb.lanl.gov) contains two pages from which searches can be made for mutations known from the literature to confer resistance to a variety of antiretroviral drugs. This database is updated annually with input provided by Dr. John Mellors and his staff at the University of Pittsburgh.

Simple search

On the simple search page, the searchable fields include gene, compound, drug class, and amino acid position. In the example shown here, the user is searching for mutations in all genes and at any positions associated with compounds that contain the "word" PNU.

Simple search

This search produces the results page illustrated below.

A tabular output shows several basic fields for 10 of the 16 mutations for the compound PNU-140690 (tipranavir). To see the next 6 records, press the Forward button. To see more detailed information about any particular mutation, click on the blue link in the first column "HIV-1 Protease", which takes you to the "record detail" layout shown on the next page.

All information in the database about this mutation is displayed in this view. There is a link to the PubMed entry for the citation associated with this mutation.

Gene: **HIV-1 Protease**   DrugClass: Protease inhibitor

Compound: PNU-140690 (tipranavir)   Synonyms: PNU-140690, tipranavir, U-140690

AAMutation: **L 10 F**   Codon mutation: CTC -> TTC

In Vitro: Y In Vivo:?   Selected or CrossResist: Selected

Comment: After 73 passages 9 other mutations accumulated to give 87 fold resistance to tipranavir

---

L. Doyon, S. Tremblay, L. Bourgon, E. Wardrop, M.G. Cordingley
Selection and characterization of HIV-1 showing reduced susceptibility to the non-peptidic protease inhibitor tipranavir
Antiviral Res **68(1)**: 27-35 (2005) Medline link: 16122817

Press the **Back** button in your browser to return to table view, or [Find]

Advanced search

The advanced search page gives you a much wider range of fields on which to search. The illustration below shows a search for protease mutations whose author list contains the name "Condra". (Fourteen records are found.) The layout and information presented on the results pages for the advanced search are identical to those in the simple search.

ADRA

ADRA, the Antiretroviral Drug Resistance Analysis site (http://www.hiv.lanl.gov/content/hiv-db/ADRA/adra2.html), can be considered both a method of sequence analysis and a search tool. Given a query sequence as input, ADRA scans the sequence to identify the presence of mutations known to confer resistance to antiretroviral drugs. ADRA includes both drugs of clinical significance and compounds that are not clinically validated.

Input ADRA requires a query sequence, either pasted in, or uploaded from a file. The user should specify whether this query is nucleotide or protein, and should select a reference sequence against which to compare the query.

Output

Mutation table ADRA produces a table, reproduced in part below, of mutations found in the query sequence ordered by amino acid position.

| Download | text file

Sequence Name	vwxyz_frame1
identified gene region(s)	Protease, RT
HXB2r positions*	56 – 479

* indicates amino acid position

● Table of mutations potentially conferring resistance(relative to HXB2r)

Protein	aa change	codon change	fold resist	cross resist	compound	record
Protease	L10I	CTC / ATC	ND	ND	MK-639 (L-735,524, indinavir)	view
Protease	L10I	CTC / ATC	ND	ND	Ro 31-8959 (saquinavir)	view
Protease	K20R	AAG / AAA	ND	ND	ABT-538 (ritonavir)	view
Protease	K20R	AAG / AAA	ND	Ro-31-8959 (8);	MK-639 (L-735,524, indinavir)	view
Protease	M36I	ATG / ATA	ND	ND	ABT-538 (ritonavir)	view
Protease	M36I	ATG / ATA	ND	ND	AG1343 (nelfinavir)	view
Protease	I54V	ATA / GTA	ND	ND	Ro 31-8959 (saquinavir)	view
Protease	I54V	ATC / GTC	ND	ND	ABT-538 (ritonavir)	view
Protease	I54V	ATC / GTC	ND	ND	MK-639 (L-735,524, indinavir)	view

The columns are self-explanatory. The last column labeled "record" provides a hyperlink to each mutation's detailed record in the resistance database. Clicking on the view link for the first mutation (L 10 I) brings up a view of this record.

Gene: **HIV-1 Protease**	DrugClass: Protease Inhibitor
Compound: **MK-639 (indinavir)**	Synonyms: MK-639, indinavir, crixivan, L735,524
AAMutation: **L 10 I**	Codon mutation: CTC -> ATC
In Vitro: N In Vivo:Y	Selected or CrossResist: Selected
Comment:	

Condra JH, Holder DJ, Schleif WA, Blahy OM, Danovich RM, Gabryelski LJ, Graham DJ, Laird D, Quintero JC, Rhodes A, Robbins HL, Roth E, Shivaprakash M, Yang T, Chodakewitz JA, Deutsch PJ, Leavitt RY, Massari FE, Mellors JW, Squires KE, Steigbigel RT, Teppler H, Emini EA
Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor
J Virol 70(12): 8270-6 (1996) Medline link: 8970946

Summary statement Following the table of mutations is a compilation of drugs or drug combinations to which the query sequence may possess a degree of resistance (shown at the top of the figure below).

Bear in mind it is inappropriate to use these results in clinical decisions about antiretroviral therapy. Only a minority of entries in the table have been clinically validated. This tool merely provides a summary of links between mutations and drugs defined in the literature.

Alignment Finally, an alignment of the user's query to the reference protein is presented. Mutations are indicated by asterisks (*).

This sequence may be associated with a degree of resistance to these drugs or drug combinations: Nelfinavir; 3TC (lamivudine); SC-52151; (+)dOTC; L-FddC; DXG; BMS 186,318; HBY 097; SKF108922; A-77003; Trevirdine; ddC; ADAMIII; TIBO R82913; AZT + 3TC; (-)dOTC; QM96521; UC-10 (645129); MKC442 (1-EBU); ddI; dOTC (BCH-10652); BILA 1906 BS; BILA 2185 BS; MK-639 (L-735,524, indinavir); Ro 31-8959 (saquinavir); 1592U89; XM323; ABT-538 (ritonavir); P9941; DMP 266 (L-743,726); L-697,661; AG1343 (nelfinavir); BILA 2011 (palinavir); (-)-FTC;

ALIGNMENT

P delineates the Protease gene region
R delineates the RT gene region
I delineates the Integrase gene region