# SCUEAL (Subtype Classification by Evolutionary ALgorithms) for HIV

*written by Sergei L Kosakovsky Pond (2006-2009)*

This directory contains reference alignments and HyPhy batch files needed to screen HIV-1 sequences and determine their subtypes. There are two modes of operation:

1. batch MPI-enabled processing of every sequence in an input file, and
2. one at-a-time serial version which screens one sequence at a time, but generates more output (including graphs in GUI enabled versions).

## Installation instructions[1]

1. SCUEAL requires a version of HyPhy built on of after Jan 1, 2009. Download and compile HYPHY from source using links and instructions at http://www.datamonk3y.org:8088/HyPhyTrac/wiki/BuildingFromSource GUI versions of HyPhy for Mac OS or Windows will be made available from the same URL.
2. Compile two versions of HyPhy on your computer using the following shell commands
   i. *Serial multithreaded build* **`$sh build.sh DEV`**
   ii. *MPI enabled build* **`$sh build.sh MPI`**
   iii. You should now have a **`HYPHYMP_DEV`** and a **`HYPHYMPI`** binaries in the **`HYPHY_Source`** directory.
3. Download and SCUEAL files from http://www.datamonk3y.org:8088/HyPhyTrac/changeset/latest/HBL/sergei/SCUEAL?old_path=/&filename=SCUEAL&format=zip
4. You should now have a **`SCUEAL`** directory with some files and directories.

| | |
|---|---|
| i. *README.pdf* | this file |
| ii. *data* | directory for reference alignments and supporting files; consult the README file inside that directory for an explanation of what each file is) |
| iii. *HBF* to | private HyPhy scripts used by SCUEAL; you should not need look at these files or to modify them. |

[1] for Linux, terminal in MacOS X or Cygwin/MinGW for Windows; for graphical interface Mac/Windows versions just download the HyPhy binary - those versions do not support MPI however.

iv. *Configs* — contains *settings.ibf* - a HyPhy 'include batch file' (.ibf) with user configurable options; see the comments in the file for an expla nation of possible settings; Edit the settings in this file as needed.

v. *TopLevel* — these primary files which carry out SCUEAL screening and pre pare reference alignments.

vi. *Samples* — test alignments.

**Screening sequences.**
In the following, **pathto** refers to the absolute UNIX filepath to a given directory, *e.g.* **pathto/SCUEAL** is the absolute path of the SCUEAL directory. Input sequences must be in frame coding sequences (with at most one frameshift) or nucleotide sequences, and represent the same genomic region as the reference sequences. The code will run if you try to screen *env* sequences using a *pol* reference alignment, but the results will clearly be nonsensical.

**Batch screening of all sequences in a FASTA file**

1. Execute the following shell command: **$mpirun -np [number of processors] pathto/HYPHYMPI BASEPATH=pathto/HYPHY_Source/ USEPATH=/dev/null pathto/SCUEAL/TopLevel/MPIScreenFASTA.bf**

2. When prompted select whether the alignment should be treated as in-frame codons or nucleotides, enter the path to the alignment to screen (either relative to **pathto/SCUEAL/TopLevel** or absolute) and the path for the resulting tab separated file. The final 'Result output option' allows you to instruct HyPhy to write out two files per input sequence into a directory of your choice (make sure the path given here ends with a slash, e.g. /home/sergei/SCUEAL/Results/). For each sequence an ID.ps and and ID.lf file will be written to the results directory (ID is replaced with the numeric position of the query in the file, e.g. the results for the 5th sequence will be written to files 5.ps and 5.lf). See the next section for the description of PostScript and .lf (HYPHY batch file save file).

3. Note: the previous step can be wrapped (in sh or bash for example) by piping, e.g **$(echo 1; echo file_in; echo file_out; echo 1) | mpirun -np [number of processors] pathto/HYPHYMPI ... [snipped]**

4. The analysis will run, displaying feedback, which node is processing which job and what subtype was identified for each sequence as they are processed, e.g.

```
[SEND] Sequence 18_CPX.CM.97.CM53379 ACC AF377959 to MPI node 9
[RECEIVE] Sequence 01_AE.TH.90.CM240 ACC U54771 from node 3 (16 alignments remaining): AE
[SEND] Sequence 19_CPX.CU.99.CU38 ACC AY588970 to MPI node 3
```

5. The resulting tab separated file will contain 6 columns (although not every sequence will have a value in each column)

| Column | Meaning |
|---|---|
| **`Index`** | Index of the sequence in the input alignment |
| **`Name`** | Sequence label from the input alignment |
| **`Subtype`** | Inferred sequence subtype<br>1. Error (e.g. **`Error: alignment failed`**) occurs when the query sequence could not be automatically aligned with the reference set, typically due to multiple frameshifts/premature stop codons<br>2. Non-recombinant, i.e. query clustered with a single sequence. The label of the subtype e.g. (**`A`**) is reported or, if the query clustered with an interior reference node, ambiguous types (e.g. **`B/D`**) may be reported.<br>3. Recombinant, either within subtype (reported as e.g. **`B inter-subtype recombinant (2 breakpoints)`**), or between-subtype, e.g. **`A1/AE,B inter-subtype recombinant`**. |
| **`Support`** | Model averaged support for the subtype assignment. Lower values (<0.9) indicate that there may be subtype assignment ambiguity and that the sequence may need to be screened with a more specialized reference alignment or using more stringent GA settings. |
| **`Recombination Support`** | Model averaged support for presence of recombination in the sequence |
| **`Intra-subtype Support`** | Model averaged support for presence of intra-subtype recombination in the sequence |
| **`Breakpoints`** | For recombinant sequences, a list of breakpoints is given. E.g. **`745(744-746);1439(1414-1464)`** - semicolon separated list of most likely breakpoint placements, together with confidence bounds (95% CI) on the location |
| **`Sequence`** | For recombinant sequences, the exact query sequence used for screening - this provides a reference for the breakpoint coordinates reported in the previous step |

### Individual screening of a sequences from a FASTA file

This option useful for more detailed screening of 'difficult sequences' or to do it interactively. To run the analysis through a GUI version of HyPhy, use the **File:Open:Open Batch File** menu option to execute **ScreenSequenceFromAlignment.bf** (located in the **TopLevel** directory). Follow the prompts to select '**Codon**' for the alignment type select the input alignment file and which sequence to screen. Try screening **12_BF.AR.99.ARMA159 ACC AF385936** from the **LANL_CRF_sample.fas** file in **Samples**. and save the resulting file to **Sample/12CRF.ps** (or any other file you like) The analysis outputs progress report to the console like so

```
Screening 12_BF.AR.99.ARMA159 ACC AF385936
Initial subtype assignment: F1
B,F1 inter-subtype recombinant
F1,B,F1 inter-subtype recombinant
```

The results are also printed to the screen (see the previous table for explanation); additional items here include the list of (if any) alternative mosaic types having at least 5% of model average support.

```
Predicted subtype                     : F1,B,F1 inter-subtype recombinant
Model averaged support                :  69.8720%
Support for recombination             : 100.0000%
Support for intra-subtype recombination:   0.0000%

There are 2 other mosaic types with model-averaged support over 5%
     Alternative subtype        :F1,F1,B,B,F1 inter-subtype recombinant
     Model averaged support     : 18.6224%
     Alternative subtype        :F1,F1,B,F1 inter-subtype recombinant
     Model averaged support     : 11.4577%
Breakpoint   1:      751bp, 95% confidence range: 747-755 bp.
Breakpoint   2:     1442bp, 95% confidence range: 1433-1451 bp.
CCTCAAATCACTCTTTGGCAACGACCCCTAGTCATAATAAAAGTAGGGGGACAGCTAAAGGAAGCTCTATTAGATACAGGAGCAGA
TGATACAGTATTAGAAGACATAAATTTGCCAGGAAAATGGAAACCAAAAATGATAGGGGGAATTGGAGGTTTTATCAAAGTAAAAC
AGTATGATAACGTACTCATAGAAATTTGTGGACACAAGGCTATAGGTACAGTGTTAATAGGACCTACACCGGTCAACATAATTGGA
AGAAATCTGTTGACTCAGCTTGGTTGCACTTTAAATTTTCCCATTAGTCCTATTGAAACTGTACCAGTAAAATTAAAGCCAGGAAT
GGATGGCCCAAAAGTTAAACAATGGCCATTGACAGAAGAAAAAATAAAAGCATTAACAGAAATATGTACAGAAATGGAAAAAGAAG
GAAAAATTTCAAAAATTGGGCCTGAAAATCCATACAATACTCCAGTATTTGCCATAAAGAAAAAAGACAGTACTAAATGGAGGAAA
TTAGTAGATTTCAGAGAACTTAATAAAAGAACTCAAGATTTTTGGGAGGTGCAATTAGGAATACCGCATCCTGCAGGGTTAAAAAA
GAAAAAATCAGTAACAGTACTAGATGTGGGGGATGCATACTTTTCAGTCCCCTTAGATGAGGCTTTCAGGAAGTACACTGCATTCA
CCATACCTTGTGTCAACAATGAGACACCAGGAACTAGGTACCAGTACAATGTGCTTCCACAGGGGTGGAAAGGATCACCAGCAATA
TTCCAAAGCAGCATGACAAAGATCCTAGAGCCTTTTAGAAAACAAAATCCGGACATAGTTATCTATCAATACATGGATGATTTGTA
TGTAGGATCTGATTTAGAAATAGGGCAGCATAGAACGAAAATAGAGGAACTAAGACAACATCTGTTAAGGTGGGGATTTACCACAC
CAGACAAAAAACATCAGAAAGAACCTCCATTCCTTTGGATGGGTTATGAACTCCATCCTGATAAATGGACGGTACAGCCTATAGTG
CTGCCAGAAAAAGACAGCTGGACTGTCAACGACATACAGAAGTTAGTGGGAAAATTAAATTGGGCAAGTCAGATTTACCCAGGGAT
TAAAGTGAAGCAATTATGTAGACTCCTTAGGGGAACCAAGGCACTAACAGAAGTAATACCACTAACAAAAGAAGCAGAGCTAGAAC
TGGCAGAAAACAGGGAAATTCTAAAAGAACCAGTACATGGAGTGTATTATGACCCATCAAAAGACTTAATAGCAGAAGTACAGAAG
CAGGGGCAAGGTCAATGGACATATCAAATCTTTCAAGAGCCATTTAAAAATCTGAAAACAGGAAAATATGCAAGAATGAGGGGGC
CCACACTAATGATGTAAAACAATTAACAGAAGCAGTGCAAAAAATAGCCACAGAGAGCATAGTAATATGGGGAAAGACTCCTAAGT
TTAAACTACCCATACTAAAAGAGACATGGGATACATGGTGGACAGAGTATTGGCAAGCCACCTGGATTCCTGAATGGGAATTTGTC
```

```
AATACCCCCCCTCTAGTAAAACTATGGTATCAGTTAGAAACAGAGCCCATAGCAGAAGCAGAAACCTTCTATGTAGATGGGGCATC
TAATAAAGAGACCAAAAAAGGAAAAGCAGGATATGTTACTGACAAAGGAAGACAAAAAGTTGTCTCCCTAACTGAAACCACAAATC
AGAAGGCTGAGTTACATGCAATTTACTTAGCTTTACAGGATTCAGGATCAGAAGTAAACATAGTAACAGACTCACAGTATGCATTA
GGAATTATTCAAGCACAACCAGATAAGAGTGAGTCAGAGTTAGTCAGTCAAATAATAGAGCAATTAATAAAAAAGGAAAAGGTCTA
CCTGTCATGGGTACCAGCACACAAGGGGATTGGAGGAAATGAACAAGTAGATAAATTAGTCAGTGCTGGGGTCAGAAAAATACTGT
TTTTAGATGGGATAGATAAGGCACAGGAGGAACATGAAAAATATCACAACAATTGGAGAGCAATGGCTAGTGATTTTAATCTGCCA
CCTGTAGTAGCAAAAGAAATAGTAGCTAGCTGTGATAAGTGTCAGCTAAAAGGGGAAGCCATGCATGGACAAGTAGATTGTAGTCC
AGGGATATGGCAATTAGACTGTACACATTTAGAAGGAAAAACTATCCTGGTAGCAGTCCATGTAGCCAGTGGGTACCTAGAAGCAG
AAGTTATCCCAGCAGAAACAGGACAAGAAACAGCCTACTTCATACTAAAGTTAGCAGGAAGATGGCCAGTAAAAACAATACATACA
GACAATGGCCCCAATTTCATCAGTGCCATGGTTAAGGCAGCCTGTTGGTGGGCAGGTATCCAACAGGAATTTGGAATTCCCTACAA
CCCCCAAAGTCAGGGAGTAGTAGAATCTATGAATAAAGAGCTAAAAAAGATCATAAGCCAGGTAAGAGATCAAGCTGAACATCTTA
AGACAGCAGTGCAAATGGCAGTATTCATCCACAATTTTAAAAGAAAAGGGGGGATTGGGGGATACAGTGCAGGGGAAAGAATAATA
GACATAATATCAACAGACATACAAACTAGAGAATTACAAAAACAAATTATAAAAATTCAAAATTTCCGGGTTTATTACAGGGACAG
CAGAGACCCAGTTTGGAAAGGACCAGCAAAGCTACTCTGGAAAGGTGAAGGGGCAGTAGTCATACAAGACGATAGTGAAATAAAGG
TAGTACCCAGAAGAAAAGCAAAGATCATTAGGGACTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAGGTAGACAGGAT
```

The resulting PostScript file contains three sections: result summary, mosaic plot and reference trees with the placed query sequence.

Result summary repeats the output previously printed to the screen, briefly stating the result of the analysis.

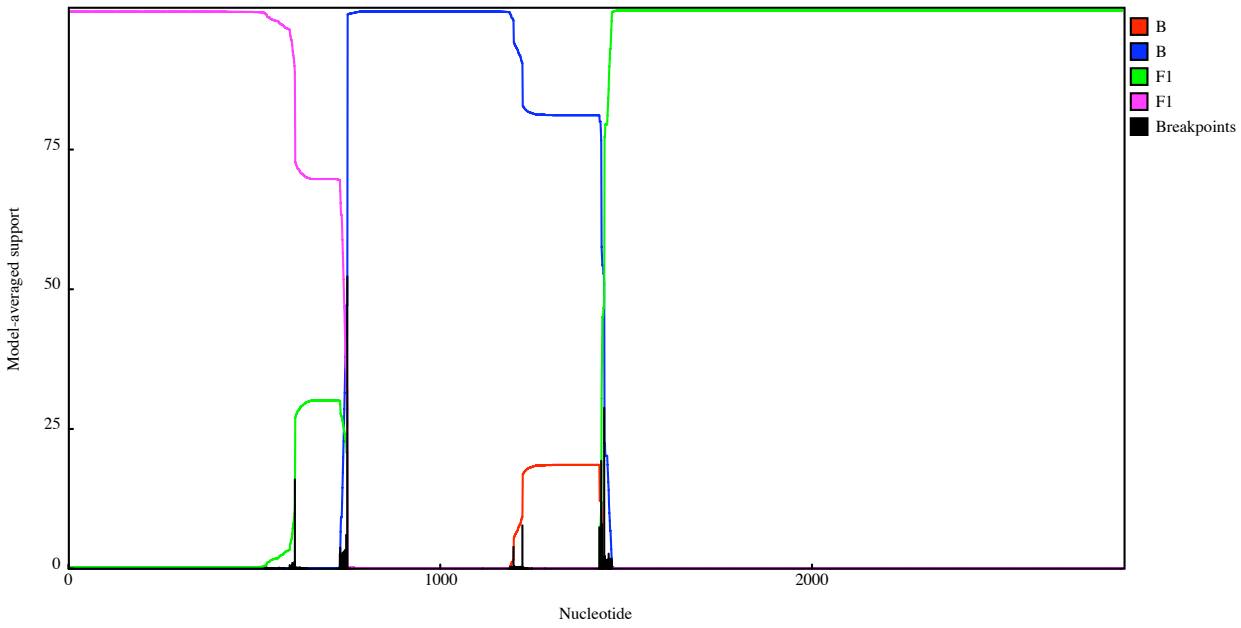
SCUEAL subtyping report for 12_BF.AR.99.ARMA159 ACC AF3...

| | |
|---|---|
| Predicted subtype | F1,B,F1 inter-subtype recombinant |
| Model averaged support | 69.8720% |
| Recombinant | 100.0000% |
| Intra-subtype recombinant | 0.0000% |

| Alternative subtype | Model averaged support |
|---|---|
| F1,F1,B,B,F1 inter-subtype recombinant | 18.6224% |
| F1,F1,B,F1 inter-subtype recombinant | 11.4577% |

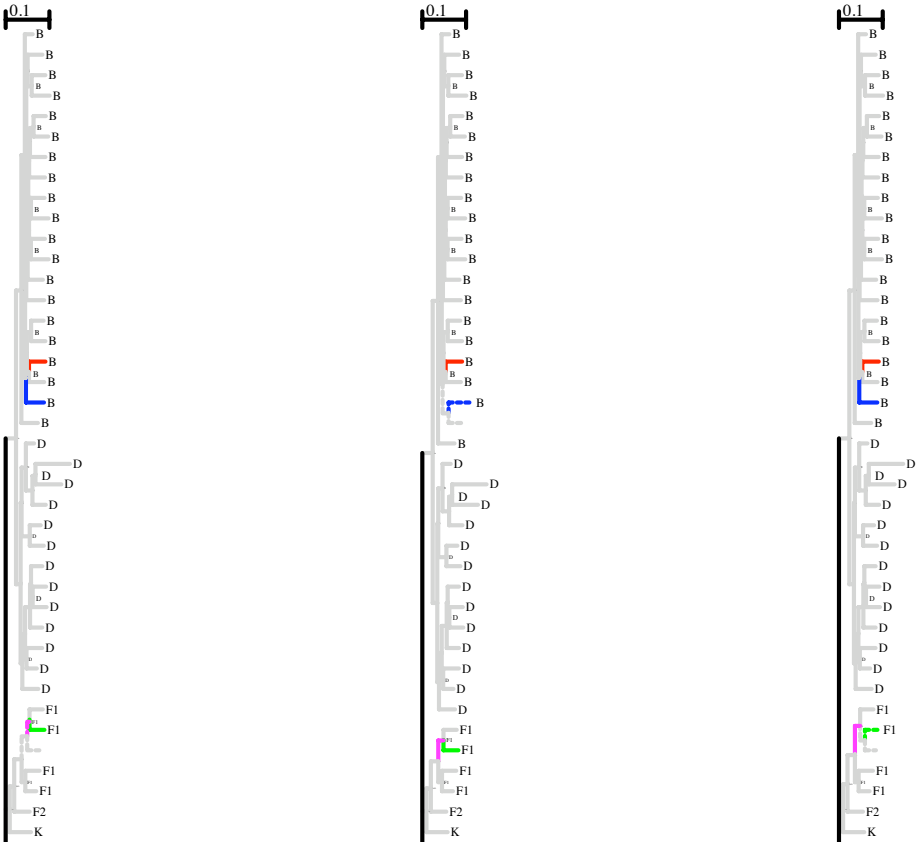
| Breakpoint locations |
|---|
| 751bp, 95% confidence range: 747-755 bp. |
| 1442bp, 95% confidence range: 1433-1451 bp. |

Mosaic plot is a graphical representation of the inferred mosaic structure. For each position on the x-axis (a site in the query sequence), model averaged support for each possible sister reference lineage is plotted (only those lineages that have 5% or greater support at one or more sites are shown). Black impulse overlay indicates the inferred location of breakpoints. Lineages are color coded and labeled by their subtype; the colors are consistent with the next section which shows their location in the reference topology.




A part of the reference topology spanning all of the lineages in the mosaic plot is rendered in the last section. The number of tree plots corresponds to the number of fragments in the most likely mosaic assignment (3 in this case: F1-B-F1); the query sequence, its sister lineage and their parent lineage are plotted with dashed lines and all leaves and internal nodes (space permitting) are labeled with their respective subtypes.

The .lf file created by SCUEAL contains a NEXUS file with the reference alignment, query sequence aligned to it and HyPhy instructions needed to construct and evaluate the phylogenetic likelihood of the best fitting mosaic model. It should be opened as a batch file in a GUI version of HyPhy to examine parameter estimates, draw trees etc.




## Preparing a reference alignment.

There are two ways to prepare a reference alignment: using all sequences from a curated alignment, or incrementally build an alignment up from a seed reference alignment.

The **`MakeReferenceAlignment.bf`** file can be used to prepare an input alignment of coding or nucleotide sequences for use in SCUEAL screening. Please note that it is assumed that there is **no recombination in any of the reference sequences**; you should run GARD (http://www.datamonkey.org/GARD) or another recombination screening tool on the reference alignment to ensure that this is the case. The input sequences can be aligned automatically by HyPhy (not recommended for gappy genes), or input as in-frame aligned sequences. It can be run via the GUI or also from the command line:

The script prompts for:

1. Whether the alignment is to be treated as in-frame codons or nucleotides

2. The location of the input file to build alignments from;

3. Whether or not to automatically codon- or nucleotide- align input sequences;

4. Whether to perform branch swapping on the provided or automatically constructed NJ tree;

5. Where to output the reference file; you can move them about later. Two files be generated - the alignment file with reference sequences and a tree, and a **`.labels`** file storing the labels for each input sequence. They should be put inside the **`data`** directory to be usable by SCUEAL;

6. How to label reference sequences for subtyping output; make sure the labels are concise and do not have punctuation or spacing in them (e.g. A, A1, F2 are good labels, F/2 is not).

Try this script on **`Samples/AG.fasta`** to prepare an A and G subtype specific pol alignment.

The **`BuildUpReferenceAlignment.bf`** file is suitable for augmenting an existing **coding** reference alignment (prepared using **`MakeReferenceAlignment.bf`**) with more sequences from the same gene/organism. Please note that this script will perform repeated alignments with the reconstructed root ancestral sequence and may not be suitable for some of the more gappy genes (e.g. env). **Importantly, this script permits the correct addition of recombinant reference sequences (e.g. HIV-1 CRFs) by representing them as several sequences - one per non-recombinant fragment, padded with indels**.

The script prompts for:

1. An existing reference alignment with at least three sequences and a corresponding .labels file. **Warning: This file will be overwritten with the augmented alignment by the script.**

2. The location of the input file to read additional sequences from;

3. Which distance measure to use to filter new sequences; the script will only attempt to add new sequences only if their genetic distance from every other sequence currently in the reference alignment is at least a given amount.

4. Distance threshold for filtering (appropriate value depends on the gene; 0.05 is good for pol);

5. For every processed sequence the user will be asked whether or not the sequence should be added to the reference alignment and what subtype should be assigned to the new sequence.