



**Karolinska
Institutet**

Sequence editing and alignments

EDCTP ENNEA training on data management

November 14-16 2011, Muhimbili hospital, Dar Es Salaam

Irene Bontell

(irene.bontell@ki.se)

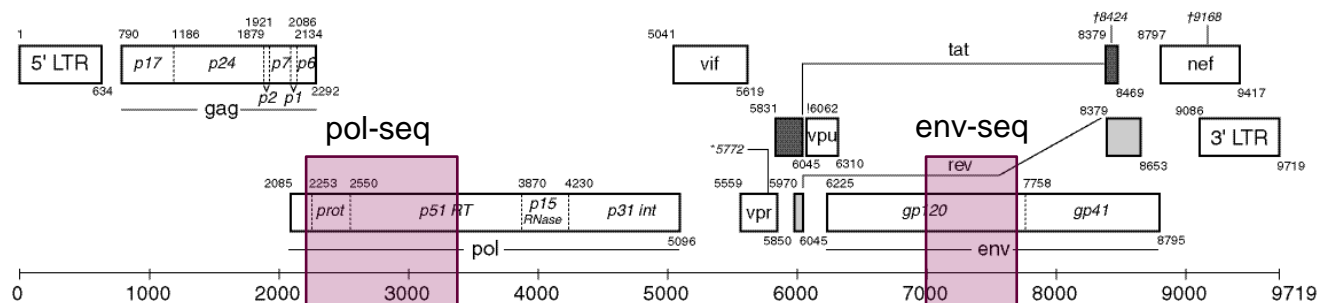
What information can we get from sequences?

For HIV common usages are:

- Subtype classification
 - Recombination analysis
 - Detection of drug resistance mutations
 - Selection of optimal antiretroviral therapy
 - Determination of co-receptor usage (V3-region of env-gene)
 - **Phylogenetic relationship between strains, time of most recent common ancestor (tMRCA)**
-

What part of the genome should I target?

- Subtype classification
 - almost any region, preferably one with many references
- Detection of recombination
 - preferably a relatively long sequence
- Determination of co-receptor usage
 - V3 region of *env*
- Detection of drug resistance mutations (transmitted or acquired)
 - for (N)NRTIs and PIs: Protease+RT in *pol*
- **Relationship between strains, time of most recent common ancestor (tMRCA)**
 - almost any region, preferably one with many references. One that is relatively conserved (like *pol*) is easier to align compared to a divergent region like *env*



Sequencing pipeline, conventional



Whole blood

→ ≥ 1 ml plasma →

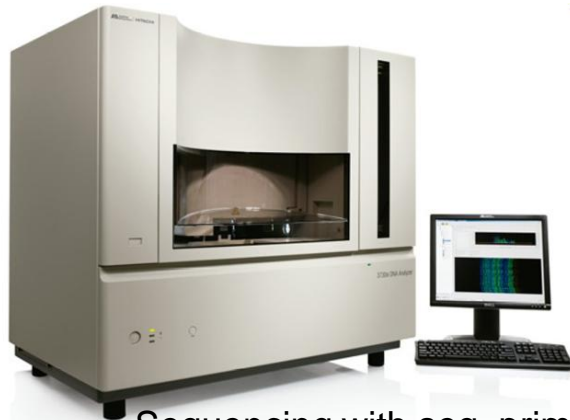


RNA-extraction

Nested PCR with specific primers



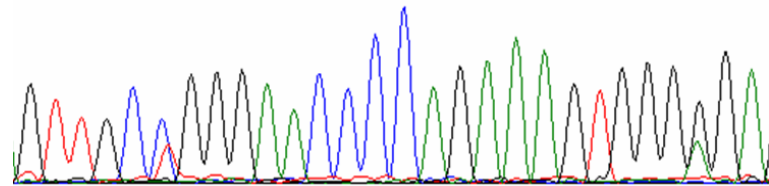
cDNA synthesis
(specific primers or random hexamers)



Sequencing with seq. primers

Output – chromatograms . Max seq length approx 1000 bp

60 70 80
G T T G C Y G G G A A C C C A G A A A G T G G R G A



Mutations in a subset of population seen as "double peaks"

Preparation of dataset

It is very important to spend the time necessary to make a good data set.

Once you have a sequence set that you trust you can go ahead with all the different analyses you want to perform.

1. Edit your sequences (aligned to reference), manual control necessary! Use BioEdit or ReCall. Quality control can be done using a LANL tool.
 2. Make a good, consistent, nomenclature and keep a file (or preferably database) with additional information about the sample and patient
 3. Perform BLAST to find the closest related sequences
 4. Download reference strains from LANL, corresponding to the same region. Use for example BioEdit to create one file with all sequences (your own + ref)
 5. Make an alignment, save in appropriate file formats (ex. FASTA and NEXUS)
 6. Check your alignment! Directly in ClustalX, or in case of a very large alignment, using Pixel
-

Sequence editing - BioEdit

BioEdit has lots of useful functions, Look through the menus and try different things!



Karolinska Institutet

The screenshot displays the BioEdit Sequence Alignment Editor interface. The main window shows two DNA sequences with their corresponding chromatograms. The top sequence is from 'Et172_ES7F_premix.ab1' and the bottom from 'Et172_ES8R_premix.ab1'. A 'view' menu is open, showing options like 'Non-editable sequence', 'Reverse Complement', and 'Positional Crosshairs'. Below the main window are two smaller windows showing detailed views of DNA sequences with various editing tools and a ruler.

Top Window: BioEdit Sequence Alignment Editor

File Edit **view** Zoom Horizontal Scale Accessory Application RNA Window Help

- Non-editable sequence
- Editable sequence
- Reverse Complement
- Positional Crosshairs
- Raw Data (DATA tags 1-4)

zania, 14-15 Nov\Practise files\Et172_ES7F_premix.ab1
File: C:\Users\Irene\Desktop\Tanzania, 14-15 Nov\Practise files\Et172_ES7F_premix.ab1

50 60 70 80 90 100 110 120 130
G A T A G T A C A G T T T A A C G A A T C T G T A C A A A T T A A C T G T A C G A G G C C C A G C A A T A A T A C A A G A A A C A G T A T A A G G A T A G G A C C A G G A C A A A C A T T

Second Window: ABI Chromatogram

Selected: none Sample: A-55948_D8 File: C:\Users\Irene\Desktop\Tanzania, 14-15 Nov\Practise files\Et172_ES8R_premix.ab1

130 140 150 160 170 180 190 200 210 220 230 240
T G A C A A A C A A T G C C A A A A T A A T A A T A G T A C A G T T T A A C G A A T C T G T A C A A A T T A A C T G T A C G A G G C C C A G C A A T A A T A C A A G A A A C A G T A T A A G G A T A G G A C C A G G A C A A A C A

Bottom Left Window: DNA sequence from C:\Users\Irene\Desktop\Tanzania

Courier New 11 B
Mode: Select / Slide Selection: 0 Position: 87

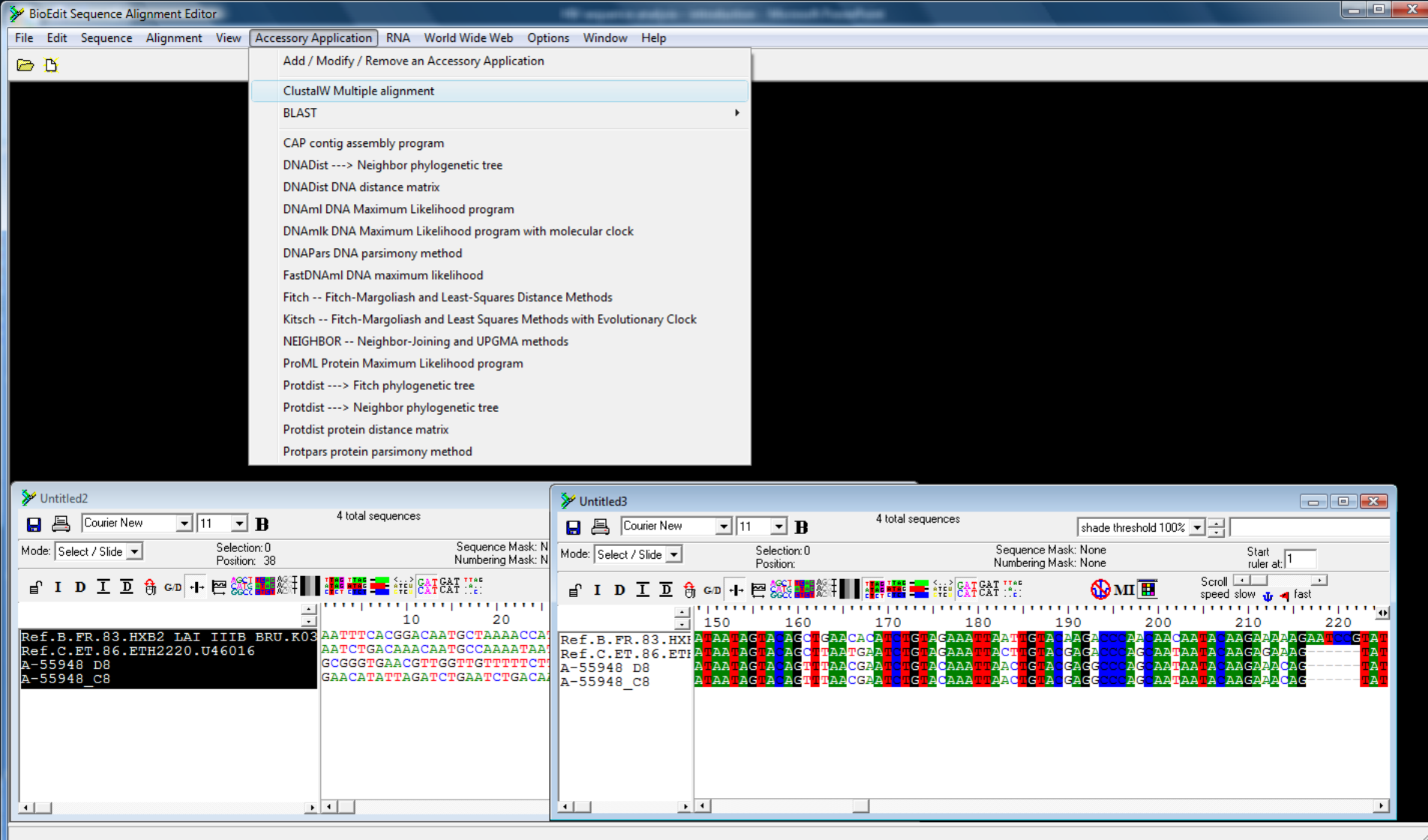
A-55948_C8 GAACATATTAGATCTGAATCT

Bottom Right Window: DNA sequence from C:\Users\Irene\Desktop\Tanzania

Courier New 11 B
1 total sequences
Mode: Select / Slide Selection: null Position: 1: A-55948_D8 50
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1
Scroll speed: slow fast

A-55948_D8 CACTGTCGATCTGATGCTCTGTGTACTTGTACTCATTGCCTCCATCTCGCAGTTCAGTAATAGTCCCTGTGATATTGAGCTACAT

Sequence editing - BioEdit



The screenshot displays the BioEdit Sequence Alignment Editor interface. The 'Accessory Application' menu is open, listing various tools such as ClustalW, BLAST, and various phylogenetic and alignment programs. Below the menu, two windows are visible: 'Untitled2' and 'Untitled3'. Both windows show sequence alignments with a ruler at the top and a sequence mask option. The 'Untitled2' window shows a sequence alignment with a ruler from 10 to 20. The 'Untitled3' window shows a sequence alignment with a ruler from 150 to 220. The 'Untitled3' window also includes a 'shade threshold 100%' option and a 'Scroll ruler at: 1' option.

Accessory Application Menu:

- Add / Modify / Remove an Accessory Application
- ClustalW Multiple alignment
- BLAST
- CAP contig assembly program
- DNADist ---> Neighbor phylogenetic tree
- DNADist DNA distance matrix
- DNAMl DNA Maximum Likelihood program
- DNAMlk DNA Maximum Likelihood program with molecular clock
- DNAPars DNA parsimony method
- FastDNAMl DNA maximum likelihood
- Fitch -- Fitch-Margoliash and Least-Squares Distance Methods
- Kitsch -- Fitch-Margoliash and Least Squares Methods with Evolutionary Clock
- NEIGHBOR -- Neighbor-Joining and UPGMA methods
- ProML Protein Maximum Likelihood program
- Protdist ---> Fitch phylogenetic tree
- Protdist ---> Neighbor phylogenetic tree
- Protdist protein distance matrix
- Protpars protein parsimony method

Untitled2 Window:

4 total sequences

Mode: Select / Slide Selection: 0 Position: 38 Sequence Mask: N Numbering Mask: N

Ref.B.FR.83.HXB2 LAI IIB BRU.K03 AATTTACGGACAATGCTAAAAACCA
Ref.C.ET.86.ETH2220.U46016 AATCTGACAAACAATGCCAAAATAA
A-55948_D8 GCGGGTGAACGTTGGTTGTTTTCCT
A-55948_C8 GAACATATTTAGATCTGAATCTGACAA

Untitled3 Window:

4 total sequences

Mode: Select / Slide Selection: 0 Position: Sequence Mask: None Numbering Mask: None Start ruler at: 1

Ref.B.FR.83.HXB2 LAI IIB BRU.K03 AATAAGGAAAGCTGAAACACATCTGACAAAATATTCATAGACCCCAACAATAAAGAAATAGAAATCTGAA
Ref.C.ET.86.ETH2220.U46016 AATAAGGAAAGCTGAAACACATCTGACAAAATATTCATAGACCCCAACAATAAAGAAATAGAAATCTGAA
A-55948_D8 AATAAGGAAAGCTGAAACACATCTGACAAAATATTCATAGACCCCAACAATAAAGAAATAGAAATCTGAA
A-55948_C8 AATAAGGAAAGCTGAAACACATCTGACAAAATATTCATAGACCCCAACAATAAAGAAATAGAAATCTGAA

Sequence editing - ReCall

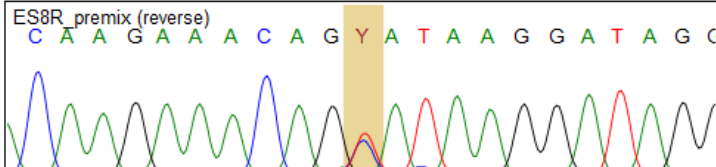
<http://pssm.cfenet.ubc.ca/home> web-based service, registration needed

Sample Et172 (id: 8143)

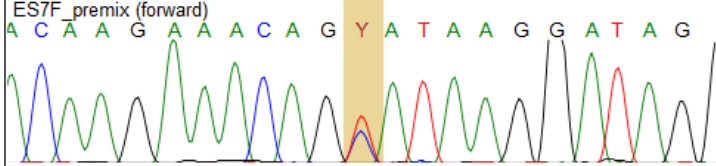
Open map

(28)	T	(29)	R	(30)	E	(31)	S	(32)	I	(33)	R	(34)	I	(35)	G	(36)	P	(37)	G	Reference Protein
	T		R		N		S		I		R		I		G		P		G	
A C A A G A G A A A G T A T A A G G A T A G G A C C A G																			Standard	
A C A A G A A A C A G Y A T A A G G A T A G G A C C A G																			Assembled	

ES8R_premix (reverse)



ES7F_premix (forward)



Job name: 2010-10-26 (id: 1343)

Upload date: 2010-10-26

Status: Passed

Mixtures: 5 (cutoff: 20.0%)

Marks: 27

"N"s: 0

Edited bases: 0

Errors: 1

• OK insert at 243 of size 3

Use the following keys to navigate:

Next marked base: right arrow

Previous marked base: left arrow

Next base: shift + right arrow

Previous base: shift + left arrow

With key locations defined in advanced settings:

Next marked key base: down arrow

Previous marked key base: up arrow

Use the following keys to make edits:

Change base: A,C,G,T,N
R,Y,K,M,S,W,B,D,H,V

Erase base: dash

Mixture compositions:

R = A/G Y = C/T K = G/T M = A/C
S = G/C W = A/T B = C/G/T D = A/G/T
H = A/C/T V = A/C/G N = A/C/GT

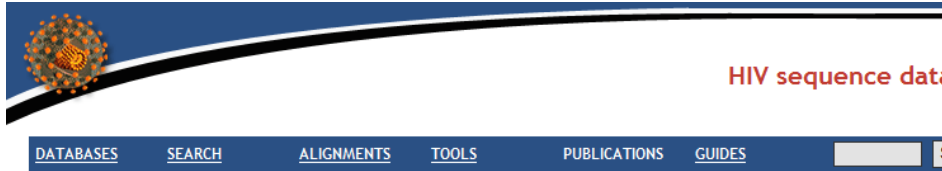
Save & Pass

Fail Sample

Exit

This program is very useful for creating a consensus sequence (much easier compared to BioEdit), but does not have the other functions of BioEdit. You need to register by e-mail, but it is worth doing if you plan on analysing a large amount of sequences

Quality control and submission of sequences



HIV sequence data

Quality Control

HIV-1 Sequence Quality Analysis

Purpose: (1) Examines sets of HIV-1 nucleotide sequences for common problems. (2) Prepares HIV-1 sequence sets, together with related data, for submission to GenBank.

Input: The tool accepts HIV-1 nucleotide sequences in [Fasta](#) format. Before using, please read the [QC/GenBank Tool Explanation](#). If you have already performed QC analyses and you only want to generate a Sequin file, you can also use the [GenBank Entry Generation](#) tool.

Input

Paste your sequence set
[Sample Input]

Upload your sequence set C:\Users\Irene\Desktop\T...

Enter a job title

Enter your e-mail address

If you find frameshifts and stop codons in your sequences, go back and look in the chromatogram to see if this really is true or if there is a mistake

Job # **23014**

Title **QC_Submission**

[NJ Tree \(all sequences\)](#)

Select

Name	Blast	RIP Subtype	Tree	Stop Codons	Frameshifts	Hypermutation	GeneCutter Result
<input type="checkbox"/> Ref.A1.RW.92.92RW008.AB253421	AB253421 RW A1 100	A1	NJ Tree	0	1	Not Detected	GeneCutter Result
<input type="checkbox"/> Ref.A1.UG.92.92UG037.AB253429	AB253429 UG A1 100	A1	NJ Tree	0	2	Not Detected	GeneCutter Result
<input type="checkbox"/> Ref.A2.CD.97.97CDKTB48.AF286238	FV536589 - - 100	A2	NJ Tree	0	2	Not Detected	GeneCutter Result
<input type="checkbox"/> Ref.A2.CM.01.01CM_1445MV.GU201516	GU201516 CM A2 100	A2	NJ Tree	0	1	Not Detected	GeneCutter Result
<input type="checkbox"/> D.TZ.2009.HM572355	HM572355 TZ D 100	D	NJ Tree	1	1	Not Detected	GeneCutter Result
<input type="checkbox"/> D.TZ.2008.HM572349	HM572349 TZ D 100	D	NJ Tree	1	1	Not Detected	GeneCutter Result
<input type="checkbox"/> D.TZ.2007.HM572350	HM572350 TZ D 100	D	NJ Tree	1	1	Not Detected	GeneCutter Result
<input type="checkbox"/> D.TZ.2009.HM572334	HM572334 TZ D 100	C,D	NJ Tree	0	1	Not Detected	GeneCutter Result
<input type="checkbox"/> Ref.C.ET.86.ETH2220.U46016	U46016 ET C 100	C	NJ Tree	0	1	Not Detected	GeneCutter Result
<input type="checkbox"/> Ref.C.ZA.04.04ZASK146.AY772699	AY772699 ZA C 100	C	NJ Tree	0	2	Not Detected	GeneCutter Result

Select

Please select *all* sequences that you want to submit. They should be submitted to GenBank as a single Sequin file.

Do not submit until you trust your sequences

Reference sequences

Under Premade Alignments one can find the subtype references. But you can also search for references in the search interface or geography search interface. Be careful when selecting your reference seq. You can't include too many (the phylogenetic analysis becomes too difficult). Include sequences that are relevant for the question you want to answer plus a couple of outgroup sequences (from a different subtype) that can be used for rooting the tree.

Programs and Tools

[Search Interface](#) retrieves HIV and SIV sequences, which can then be aligned and used to build trees

[Geography Search Interface](#) retrieves HIV sequences based on geographical distribution

[Tools for working with sequences](#) lists all our online tools, organized by function

Alignments

[HIV Premade Alignments](#) includes Consensus and Ancestral Sequences, Subtype Reference Alignments, and Complete Alignments

Alignment type: Subtype reference

Year *: 2010

Organism: HIV-1/SIVcpz

Region: Pre-defined region of the genome User-defined range

Start: 2253 End: 3263 (Coordinates: HIV1-[HXB2](#), [Mac239](#))

Subtype: All M Group (A-K + Recombinants)

DNA/Protein: DNA

Format: Fasta



HIV Sequence Alignment

Alignment type: Subtype reference

Year: 2010

Organism: HIV-1/SIVcpz

Region: User-defined range: 2253-3263 (Coordinates in the alignment: 3690)

Subtype: All M Group (A-K + Recombinants)

DNA/Protein: DNA

Format: FASTA

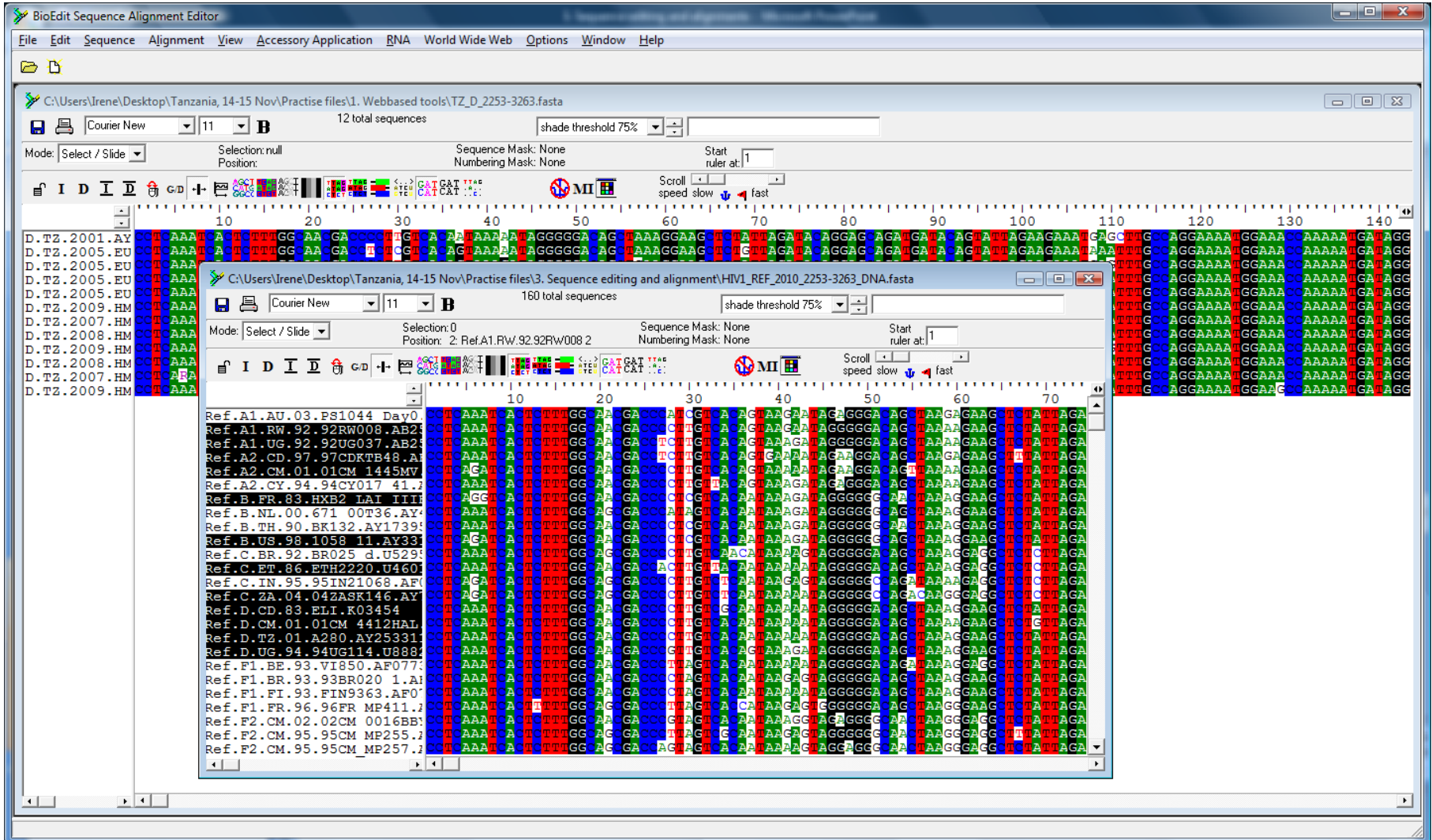
Number of sequences: 160

[Download this alignment](#)

```
>Ref . A1 . AU . 03 . PS1044_Day0 . DQ676872
CCTCAAATCACTCTTTGGCAACGACCCATCGTCACAGTAAGAATAGAGGG
ACAGCTAAGAGAAGCTCTATTAGATACAGGAGCAGACGATACAGTATTAG
AAGACATAGATTTGCCAGGAAAAATGGAAAACCAAGAATGATAGGGGAAATT
GGAGGCTTCATCAAGGTAACAGTATGATCAGATATCTATAGAAATTTG
TGGAAAAAGAGCTATAGGTACAGTATTAGTAGGACCTACACCTGTCAACA
```

Reference sequences

Merging of the file with your sequences and the references you select can easily be done in BioEdit. (Just select the references you want to include from your downloaded FASTA-file and copy and paste into the other file)



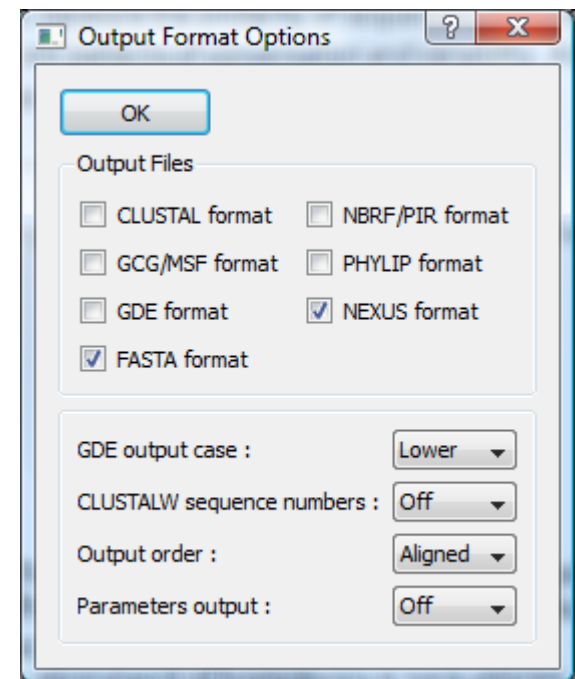
The screenshot displays the BioEdit Sequence Alignment Editor interface. The main window shows a sequence alignment of 12 total sequences from a file named 'Tanzania, 14-15 Nov\Practise files\1. Webbased tools\TZ_D_2253-3263.fasta'. The alignment is displayed in a grid format with columns numbered from 10 to 140. The sequences are color-coded by nucleotide: A (green), C (blue), G (red), and T (black). A secondary window is open, showing a sequence alignment of 160 total sequences from a file named '3. Sequence editing and alignment\HIV1_REF_2010_2253-3263_DNA.fasta'. This window also displays a grid of sequences with columns numbered from 10 to 70. The interface includes a menu bar (File, Edit, Sequence, Alignment, View, Accessory Application, RNA, World Wide Web, Options, Window, Help), a toolbar with various editing and alignment tools, and a status bar at the bottom.

Alignment ClustalX

<http://www.ebi.ac.uk/Tools/clustalx2/index.html>

After you have edited your own sequences and selected and downloaded the sequences you want to include as references it is time to make the alignment

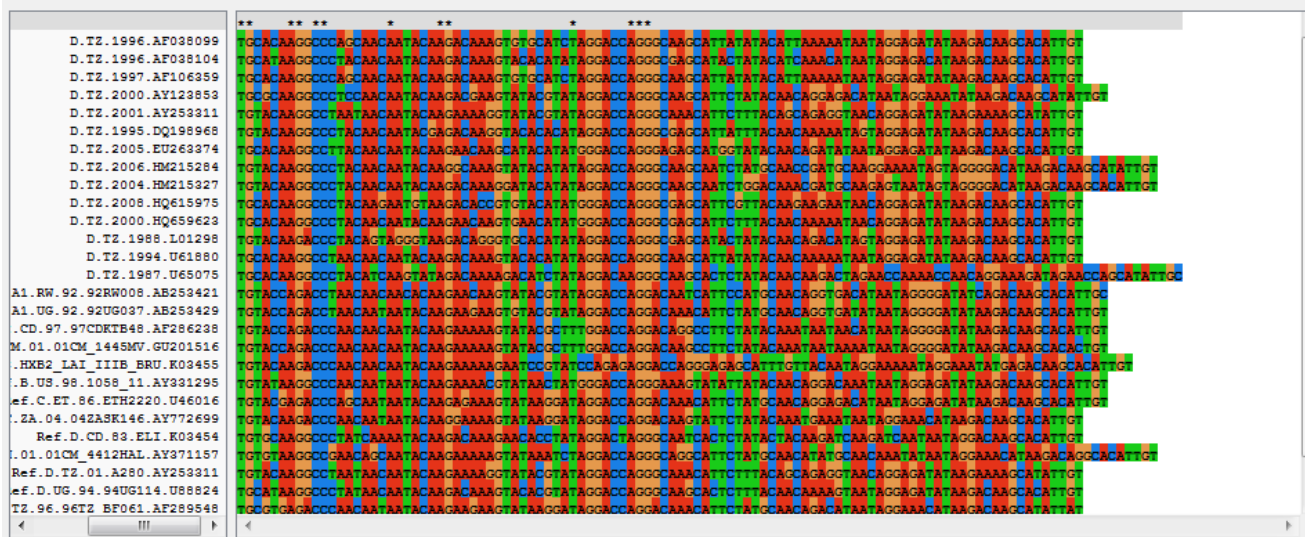
1. Make a FASTA-file with all the sequences to be included in the alignment (in BioEdit).
2. Make sure all sequences are in frame and cover exactly the same region
3. Load the FASTA file in ClustalX
4. Change the Output format options (under the Alignment menu) to FASTA and NEXUS
5. Run complete alignment
6. Manual inspection of alignment!



Pol – conserved and easy to align

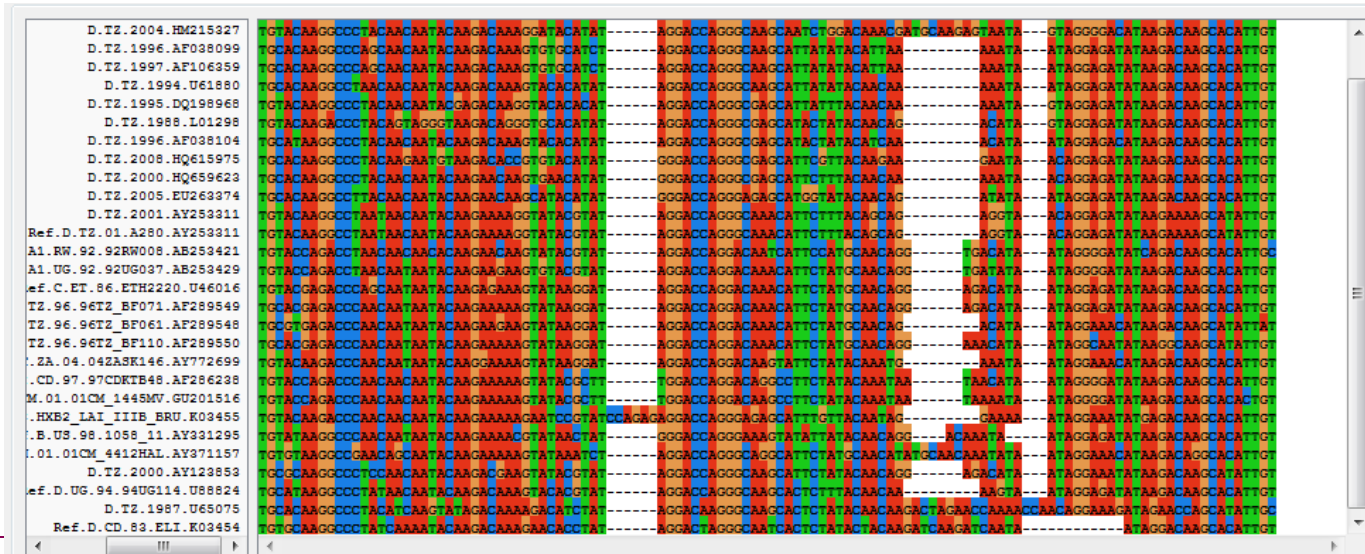


Env – less conserved, sometimes difficult to align



Sequences loaded
in Clustal before
alignment – no gaps

This alignment looks ok, but sometimes there are regions that are very difficult to align. If you can't get a good alignment it is **better to remove problematic regions** from the phylogenetic analysis. There is enough information in the rest of the sequence.



GapStrip/GapSqueeze

<http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html>



Gap Strip/Squeeze v 2.1.0

Purpose: To delete aligned columns that contain a chosen percentage of gaps or other characters.

Details: Set the gap tolerance to any value between 0% and 100%. A value of 0% will cause columns to be deleted if they contain only a single gap ("gap stripping"), while a value of 100% will delete only columns that are entirely gaps ("gap squeezing"). See [Explanation](#) file for additional details.

Input

Paste your alignment

[\[Sample Input\]](#)

Or upload your alignment

Options

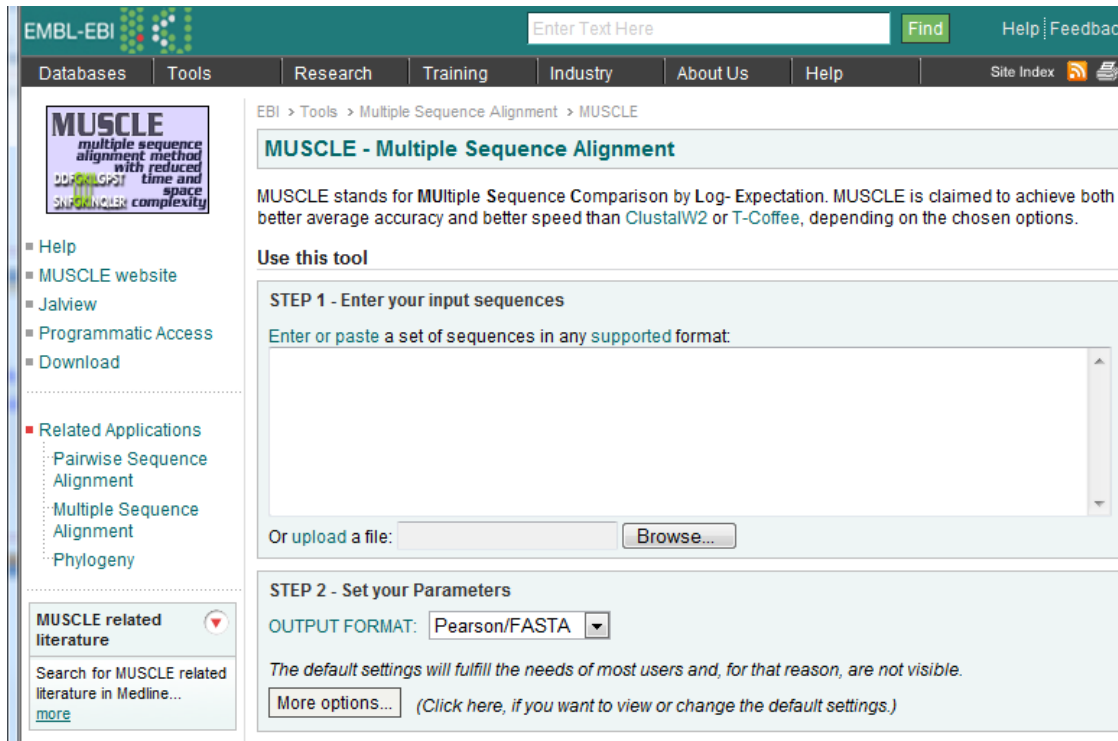
[Gap character\(s\)](#)

[Gap tolerance](#) %

Other alignment tools



CLUSTAL can be very slow for large sets.

- MUSCLE <http://www.ebi.ac.uk/Tools/msa/muscle/>
- MAFFT <http://www.ebi.ac.uk/Tools/mafft/>
- T-COFFEE <http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee.cgi/index.cgi>



The screenshot shows the EMBL-EBI website interface for the MUSCLE tool. The top navigation bar includes 'Databases', 'Tools', 'Research', 'Training', 'Industry', 'About Us', and 'Help'. The main content area is titled 'MUSCLE - Multiple Sequence Alignment' and provides a description of the tool, its advantages over ClustalW2 and T-Coffee, and instructions on how to use it. The interface is divided into two main steps: 'STEP 1 - Enter your input sequences' and 'STEP 2 - Set your Parameters'. Step 1 includes a text input field for sequences and a 'Browse...' button for file uploads. Step 2 includes a dropdown menu for 'OUTPUT FORMAT' set to 'Pearson/FASTA' and a 'More options...' button. A sidebar on the left contains navigation links for 'Help', 'MUSCLE website', 'Jalview', 'Programmatic Access', 'Download', 'Related Applications', and 'MUSCLE related literature'.

EMBL-EBI [Help](#) [Feedback](#)

[Databases](#) [Tools](#) [Research](#) [Training](#) [Industry](#) [About Us](#) [Help](#) [Site Index](#)  

EBI > Tools > Multiple Sequence Alignment > MUSCLE

MUSCLE - Multiple Sequence Alignment

MUSCLE stands for **M**ultiple Sequence Comparison by Log-Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed than *ClustalW2* or *T-Coffee*, depending on the chosen options.

Use this tool

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

Or upload a file:

STEP 2 - Set your Parameters

OUTPUT FORMAT:

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)


MUSCLE
multiple sequence
alignment method
with reduced
time and
space
complexity

- [Help](#)
- [MUSCLE website](#)
- [Jalview](#)
- [Programmatic Access](#)
- [Download](#)

.....

- [Related Applications](#)
 - Pairwise Sequence Alignment
 - Multiple Sequence Alignment
 - Phylogeny

.....

MUSCLE related literature 

Search for MUSCLE related literature in Medline... [more](#)

Alignment viewing tool: Pixel

(in the LANL tools menu)

Pixel

Show an image of an alignment with 1 pixel per residue

Purpose: This tool generates a PNG image of an alignment using 1 or more colored pixel(s) for each residue. This way, errors in large alignments can be quickly spotted. We provided an [example](#) of how to use the tool.

Input

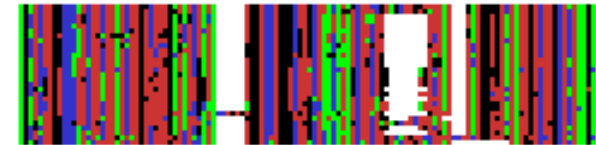
Paste your alignment here
[\[Sample Input\]](#)

or upload your file C:\Users\Irene\Desktop\Tanzania, 14-15 Nov\

Options

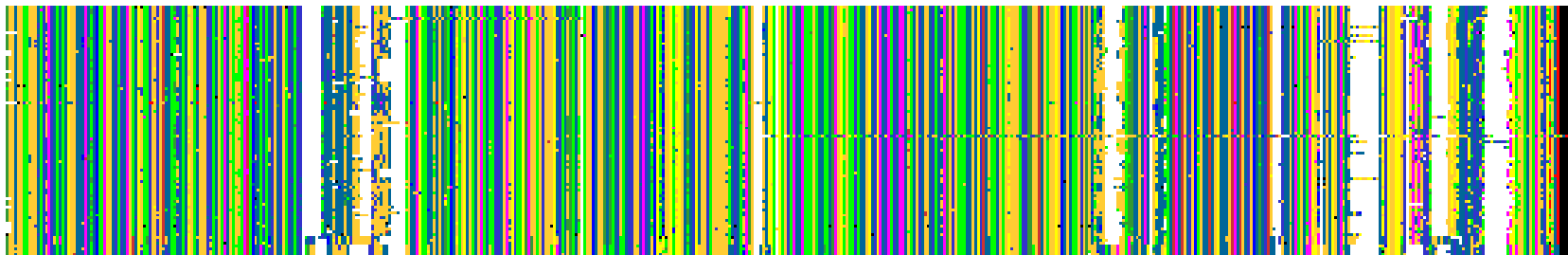
Choose sequence type Guess (by composition)
 Amino-acid
 DNA

Scale (each alignment residue will be a square of this many pixels per side)



Our V3 seq in Pixel

Example showing a misalignment



Alignment viewing tool: Pixel



Our Pol seq in Pixel

Once you are confident that you have a good alignment it is time to move on to the phylogenetic analysis!