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)

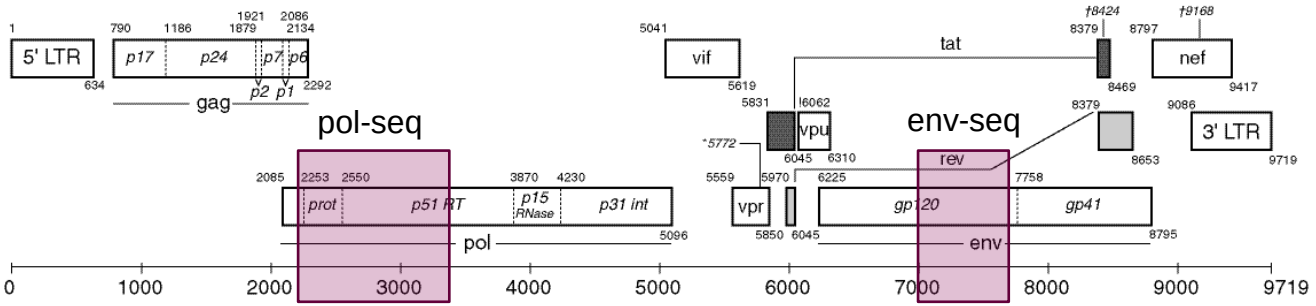
What information can we get from sequences?

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

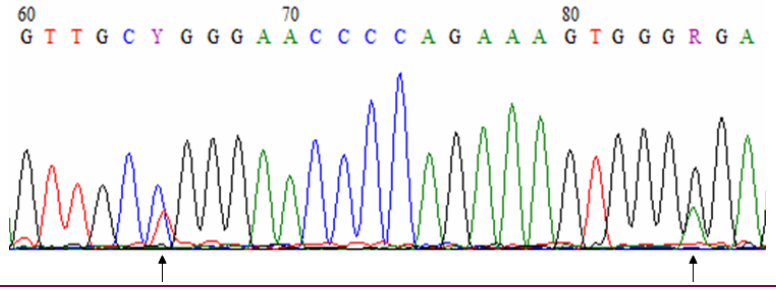
≥ 1 ml plasma

cDNA synthesis
(specific primers or random hexamers)

Nested PCR with specific primers

Sequencing with seq. primers

Output – chromatograms . Max seq length approx 1000 bp



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.

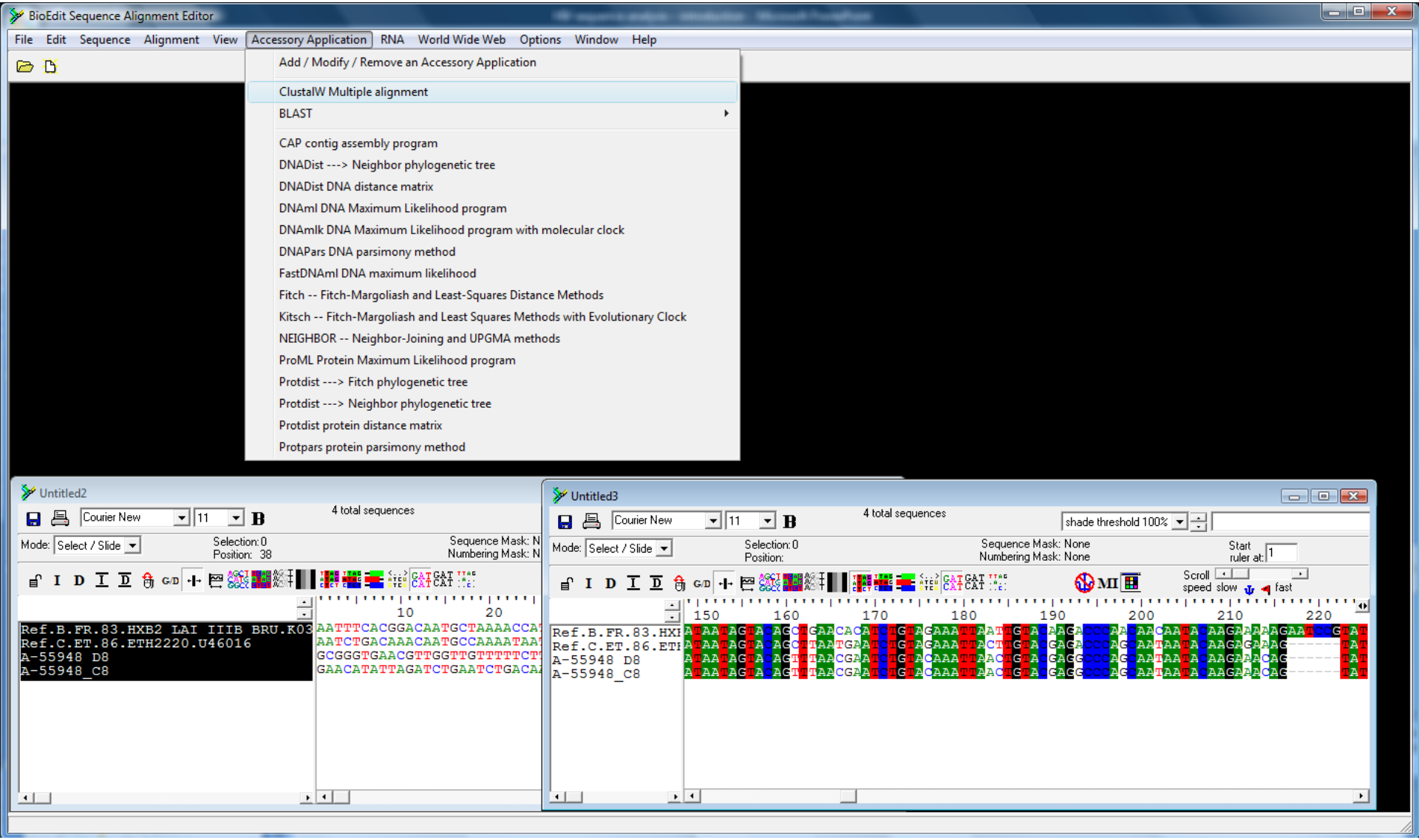
1. Edit your sequences (aligned to reference), manual control necessary! Use BioEdit or ReCall. Quality control can be done using a LANL tool.
2. Make a good, consistent, nomenclature and keep a file (or preferably database) with additional information about the sample and patient
3. Perform BLAST to 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

Sequence editing - BioEdit

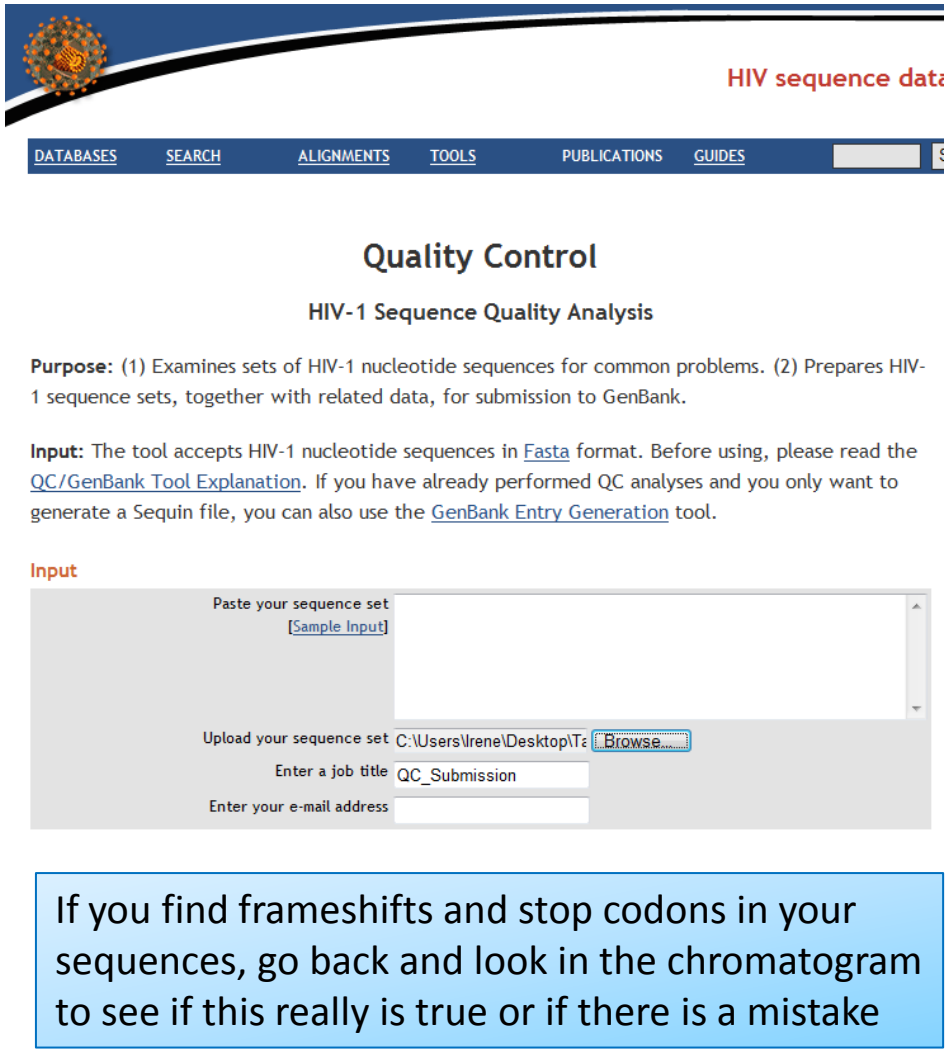


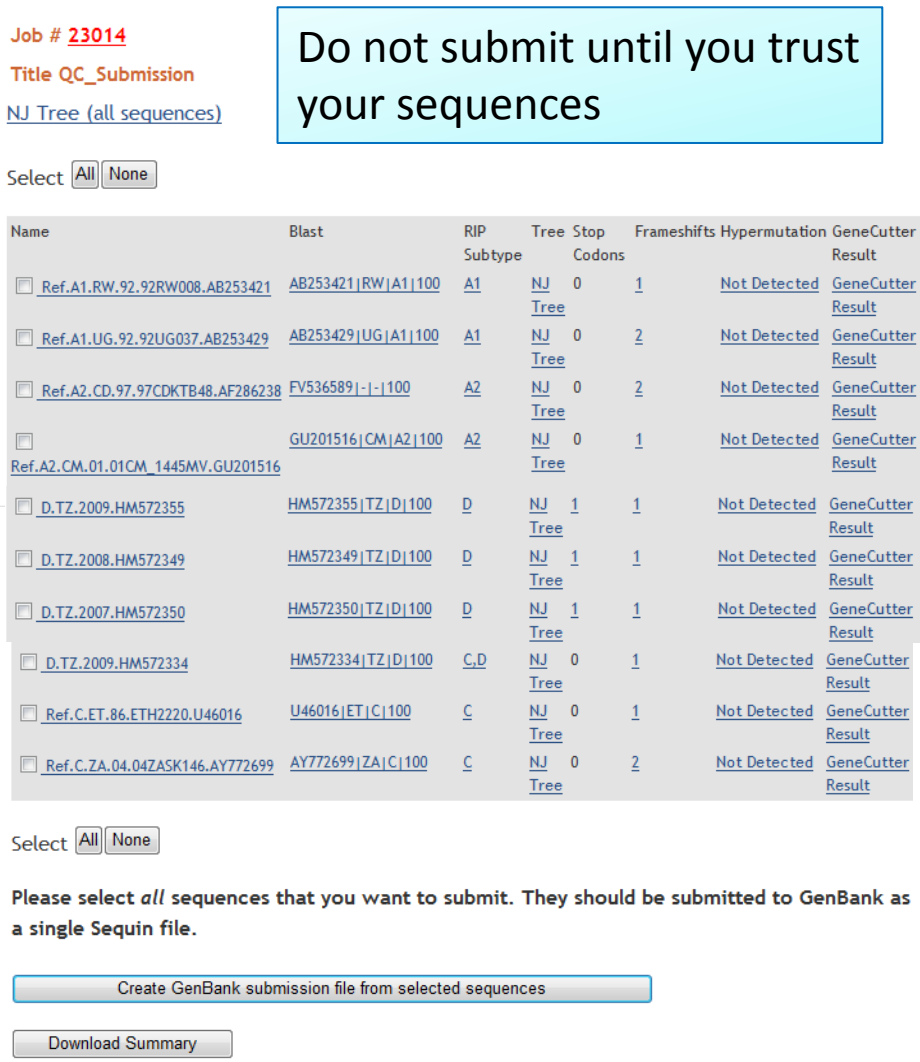
Sequence editing - ReCall

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

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



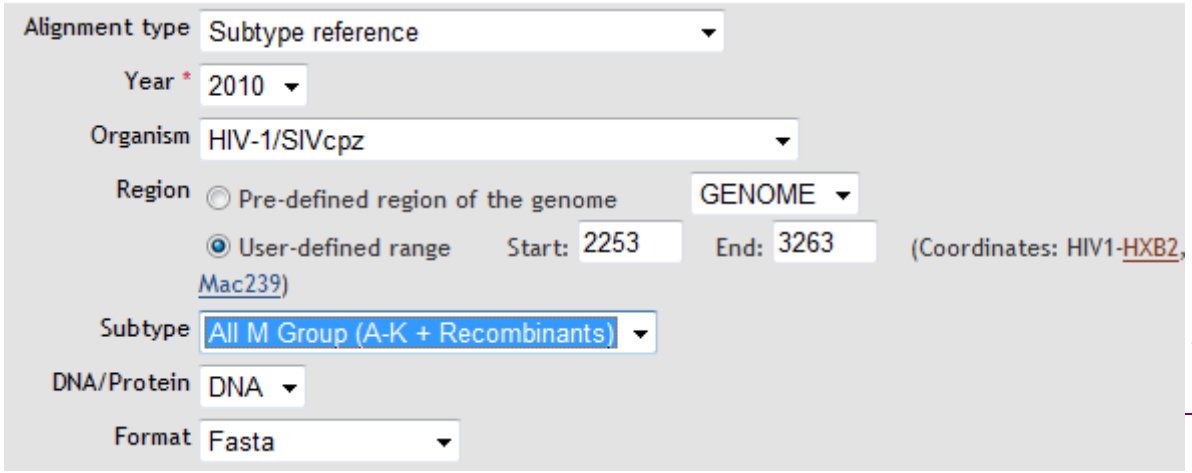


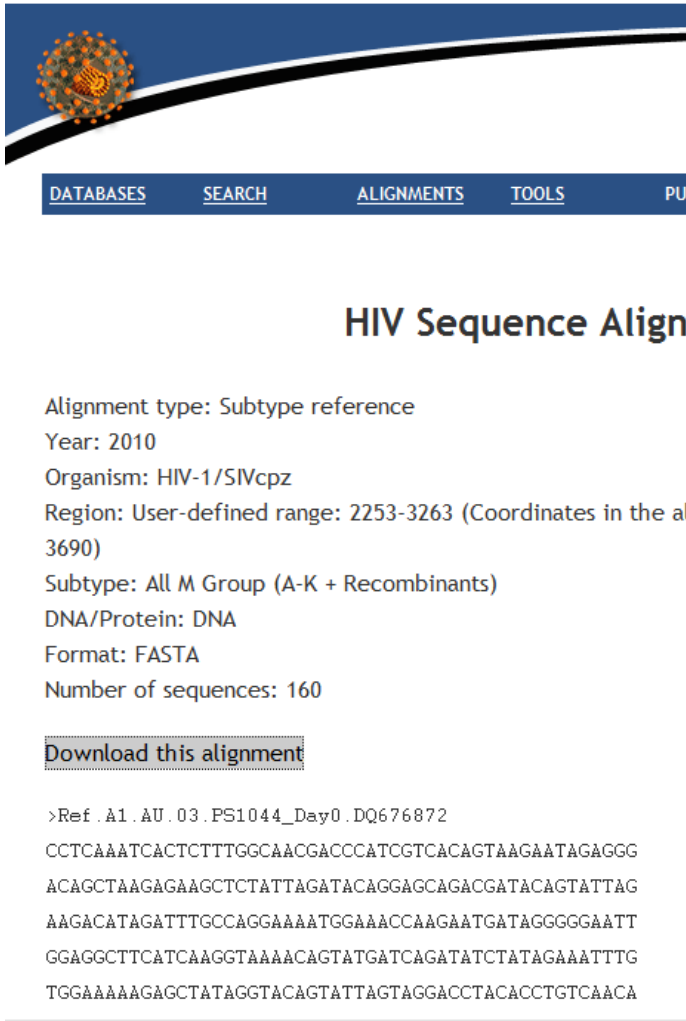
Do not submit until you trust your sequences

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

Reference sequences

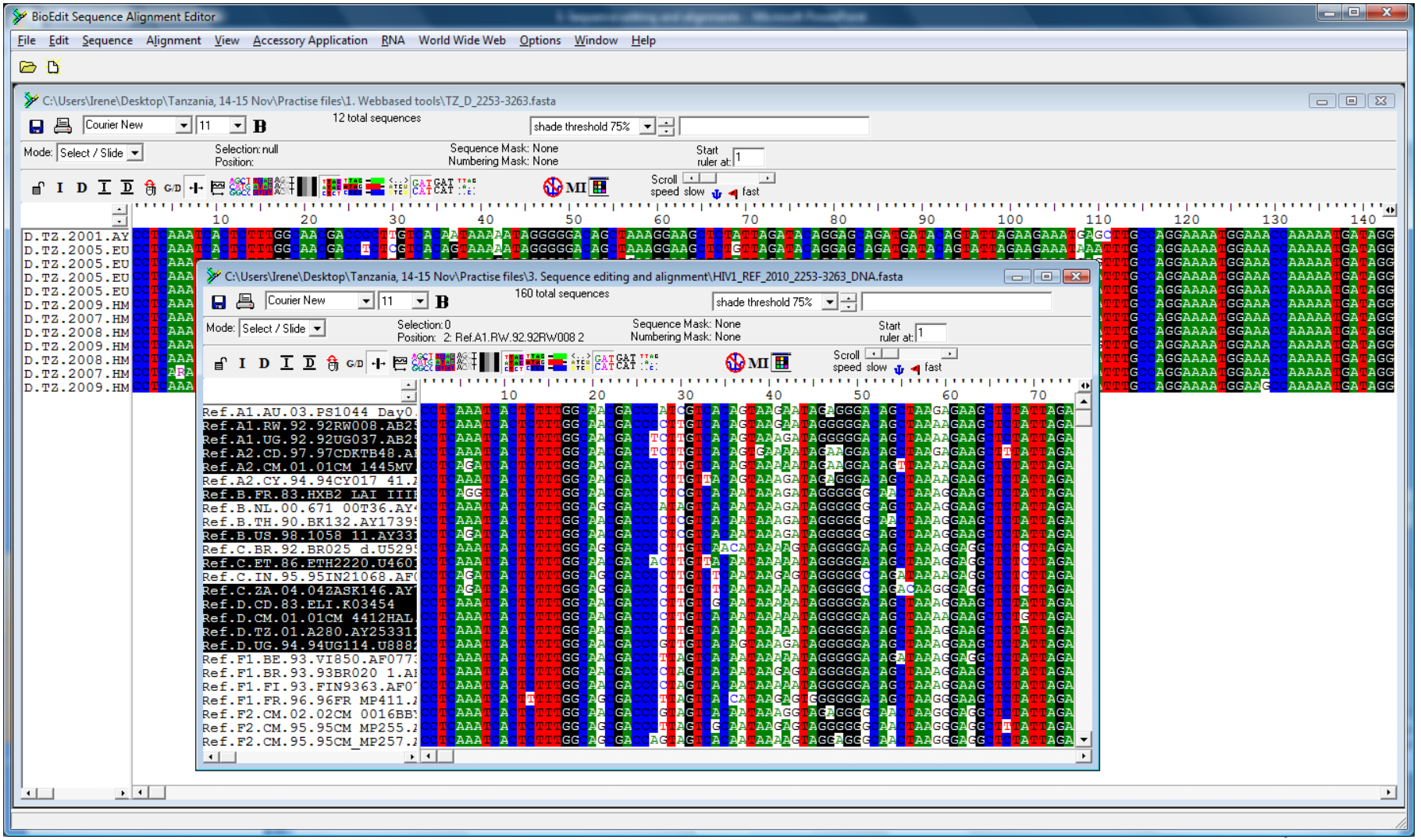
Under Premade Alignments one can find the subtype references. But you can also search for references in the search interface or geography search interface. Be careful when selecting your reference seq. You can't include too many (the phylogenetic analysis becomes too difficult. Include sequences that 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.





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)

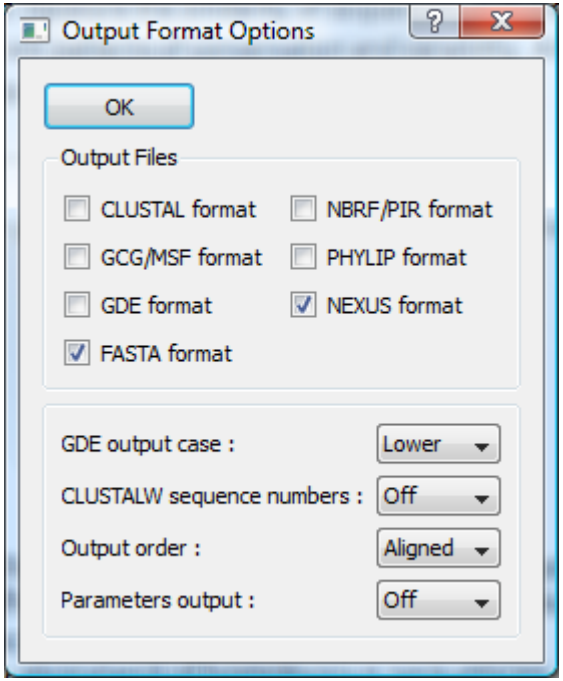


Alignment ClustalX

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

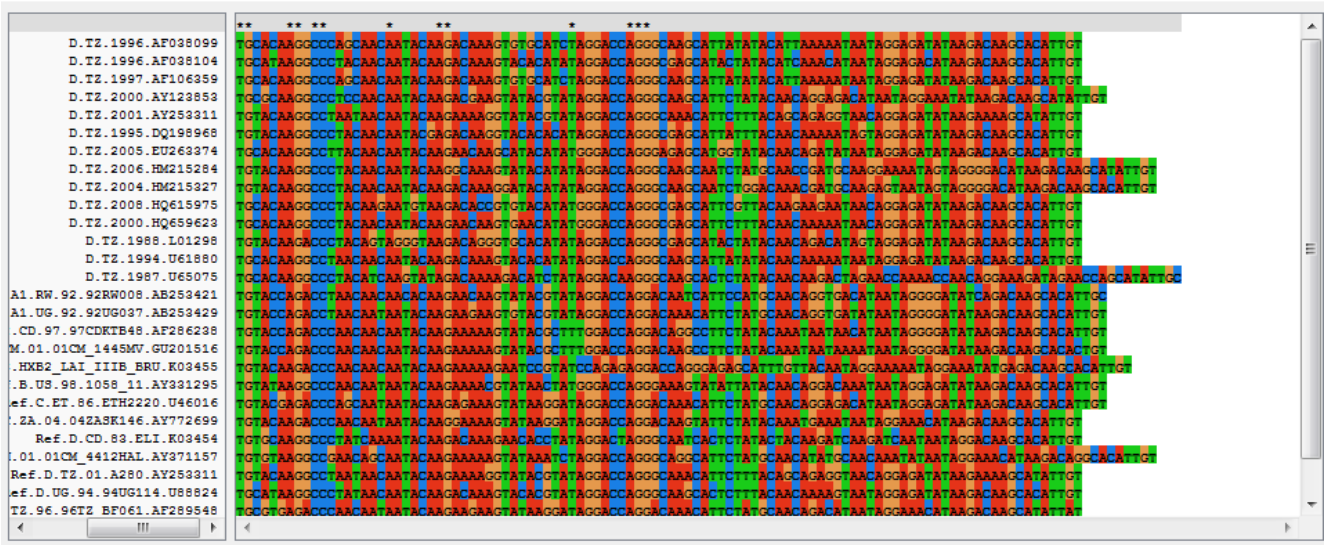
After you have edited your own sequences and selected and downloaded the sequences you want to include as references it is time to make the alignment

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!



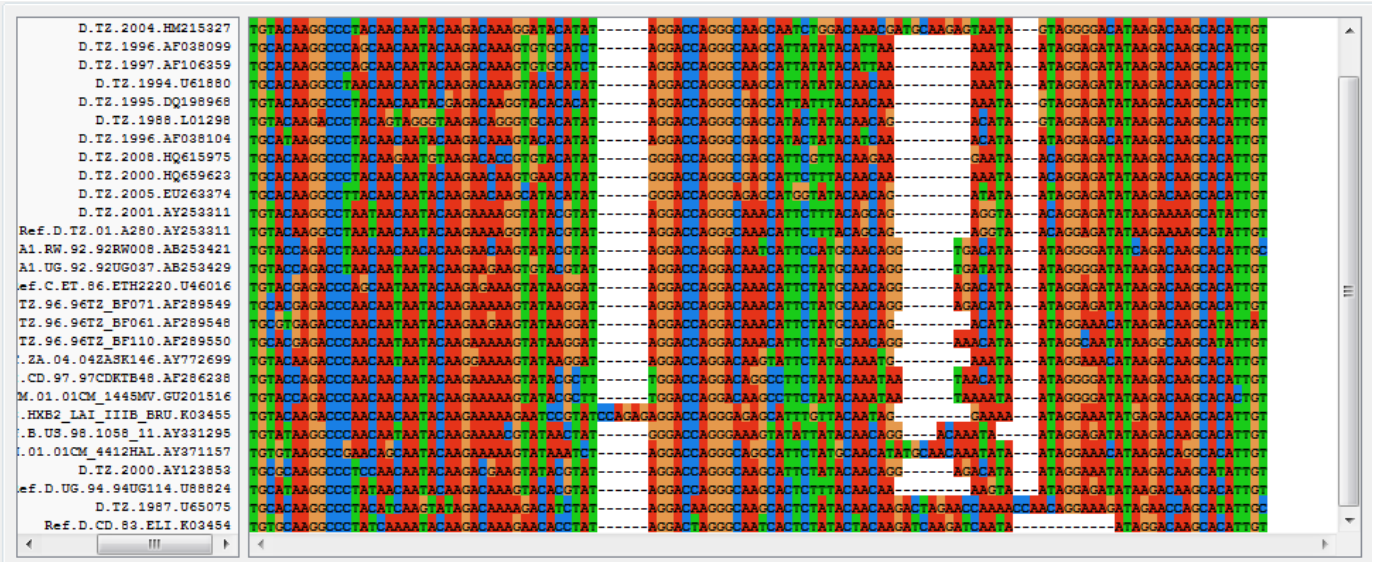
Pol – conserved and easy to align

Env – less conserved, sometimes difficult to align



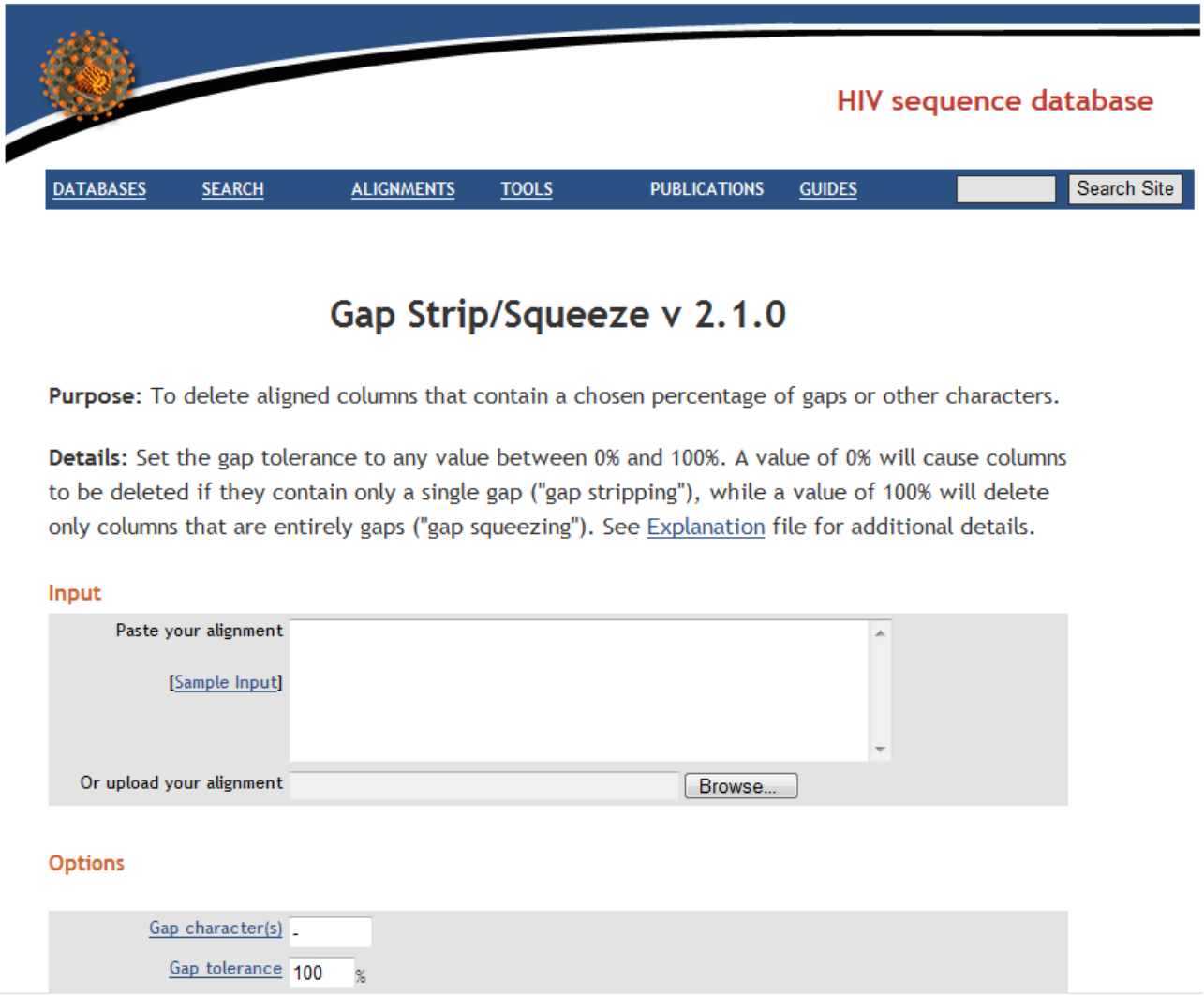
Sequences loaded in Clustal before alignment – no gaps

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

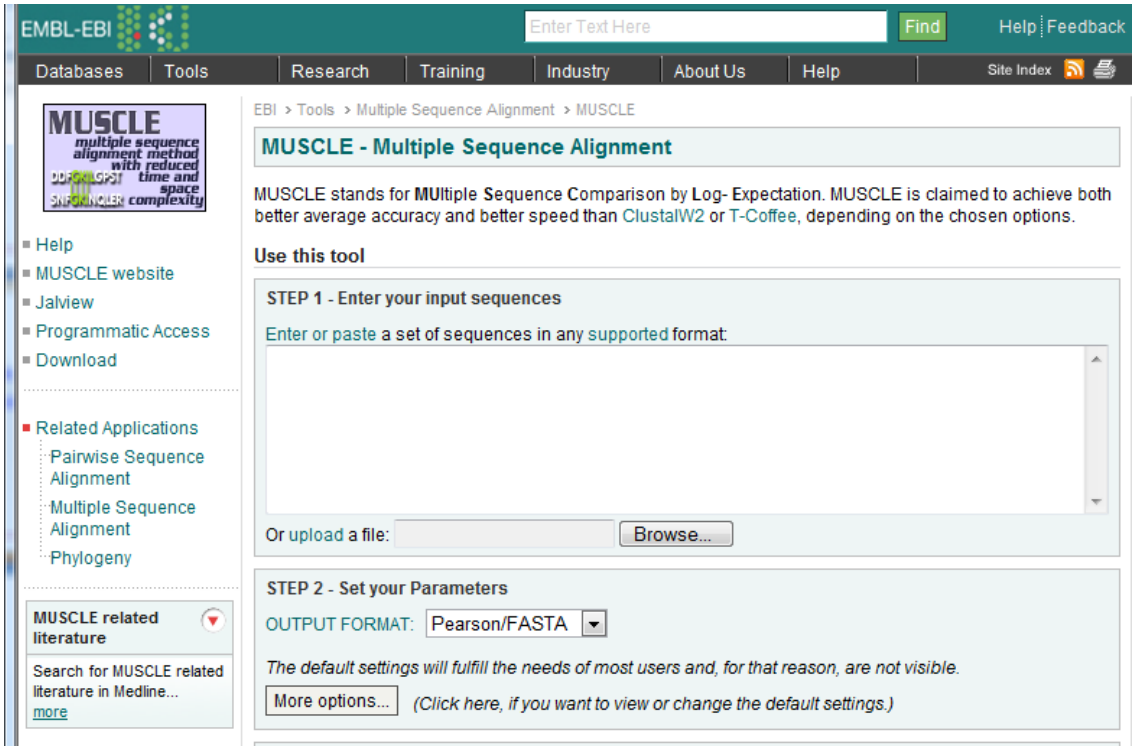
http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html



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



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]

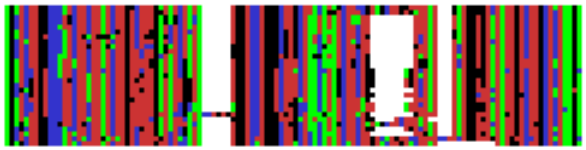
or upload your file C:\Users\Irene\Desktop\Tanzania, 14-15 Nov\P Browse...

Options

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

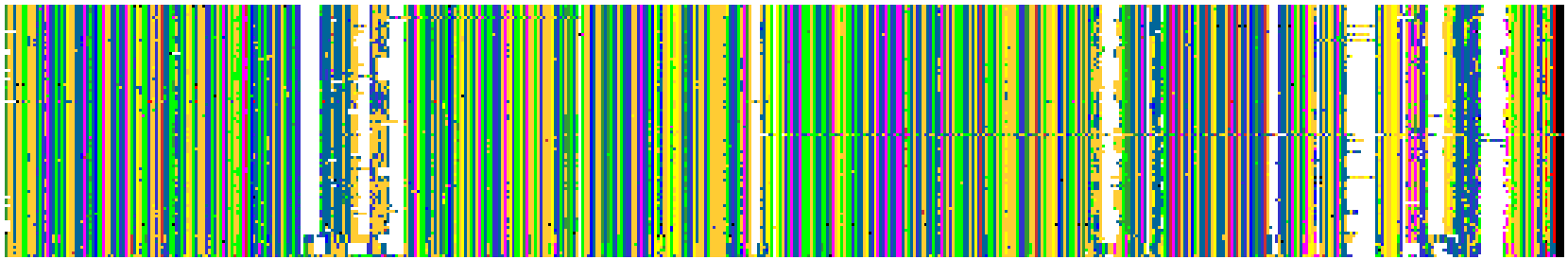
Scale 2 (each alignment residue will be a square of this many pixels per side)

Submit Reset

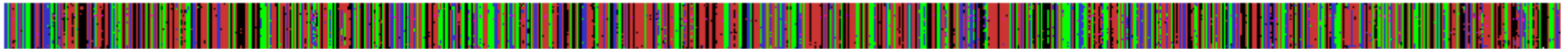


Our V3 seq in Pixel

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!