

ART-A software guide

version 0.1

Daniel Struck

Table of Contents

1	Background information.....	3
1.1	About.....	3
1.2	Features.....	3
1.3	Open-source modules.....	3
2	Installation.....	4
2.1	Running the software.....	4
2.1.1	On Windows.....	4
2.1.2	On Linux.....	4
3	Getting started.....	5
4	Importing ABI traces.....	6
4.1	Base-calling and generation of the consensus sequence.....	7
4.2	Base-call overview.....	7
4.3	Trace visualization.....	8
4.3.1	Zoom slider.....	9
4.4	Trimming.....	10
4.4.1	Automatic trimming.....	10
4.4.2	Manual trimming.....	11
5	Project view.....	13
5.1	Subtyping.....	14
6	Preferences.....	15
6.1	General preferences.....	15
6.2	Base-calling preferences.....	16
6.3	Contamination control preferences.....	17
7	Advanced tools.....	17
7.1	Contamination control.....	17
7.2	Access to data.....	17
8	Definitions.....	18
8.1	Acknowledgements.....	18

1 Background information

1.1 About

The ART-A software is a free software and open-source software for importing and analysing ab1 files from an ABI sequencer or sequences in fasta format. The base-calling capabilities are provided by the TraceTuner software.

1.2 Features

- The ART-A software can be run on a PC with a Windows or Linux operating system.
- The software is written in the Java language. Either a local Java installation (minimum version 1.6) or the Java bundled with the application can be used.
- The software can be run from a USB stick.
- Calculation intensive tasks are multi-threaded. This allows the software to scale with big datasets by providing the necessary multi-tasking capable computer platform.
- The application can run in stand-alone or network mode (server/client).
- The virus (HIV-1,HIV-2,HCV) and genes of imported sequences are automatically identified.
- Automatic contamination control.
- Automatic detection of mixed bases, ambiguities, frame-shifts, STOP codons, indels.
- Automatic list of amino acid mutations with respect to HXB2 reference strain.
- Export of mutation list to Excel.
- Export of sequences into FASTA format.
- Export of the project to VircoNET.
- Coloured and zoomable trace visualisation
- Automatic and manual trimming option available.
- All data is stored in a database.
- Server/Client mode allows to centralize data storage in the local network.
- Direct access to the database via a web browser. Queries can be formulated in the SQL language.

1.3 Open-source modules

The following open-source software packages were used to compile the ART-A software:

1. TraceTuner was used for base-calling, assigning quality values and assembly

of the sequences. TraceTuner is available under „<http://sourceforge.net/projects/tracetuner>“.

2. H2, the Java SQL database. Available under „<http://www.h2database.com>“.
3. BioJava is an open-source project dedicated to providing a Java framework for processing biological data. It provides analytical and statistical routines, parsers for common file formats and allows the manipulation of sequences and 3D structures. The goal of the BioJava project is to facilitate rapid application development for bioinformatics. In this case the modules to visualize the chromatograms have been used. BioJava is available under „<http://biojava.org>“.
4. JAligner, an open source Java implementation of the Smith-Waterman algorithm with Gotoh's improvement for biological local pairwise sequence alignment using the affine gap penalty model. It is available under „<http://jaligner.sourceforge.net/>“.
5. FigTree is designed as a graphical viewer of phylogenetic trees and as a program for producing publication-ready figures. It is available under „<http://tree.bio.ed.ac.uk/software/figtree>“.
6. JEBL, a Java library for evolutionary biology and bioinformatics, including objects representing biomolecular sequences, multiple sequence alignments and phylogenetic trees. Available under „<http://sourceforge.net/projects/jeb1>“.
7. Apache POI, Java API to access Microsoft format files. Used to export the mutations in an Excel sheet. Available under „<http://poi.apache.org>“.
8. Commons-Math: the Apache commons mathematics library. Available under „<http://commons.apache.org/math>“.

2 Installation

The latest version of the software is available at the website <http://...> (not yet in place).

Unpack the downloaded zip file either on your hard drive or on a USB stick.

2.1 Running the software

The software needs a Java runtime environment (version 1.6 at least). Java can be downloaded freely at <http://java.com>.

2.1.1 On Windows

Start the application either by launching “launch.exe” to use you installed Java runtime or “launch_bundle.exe” to use the Java bindled in the zip file.

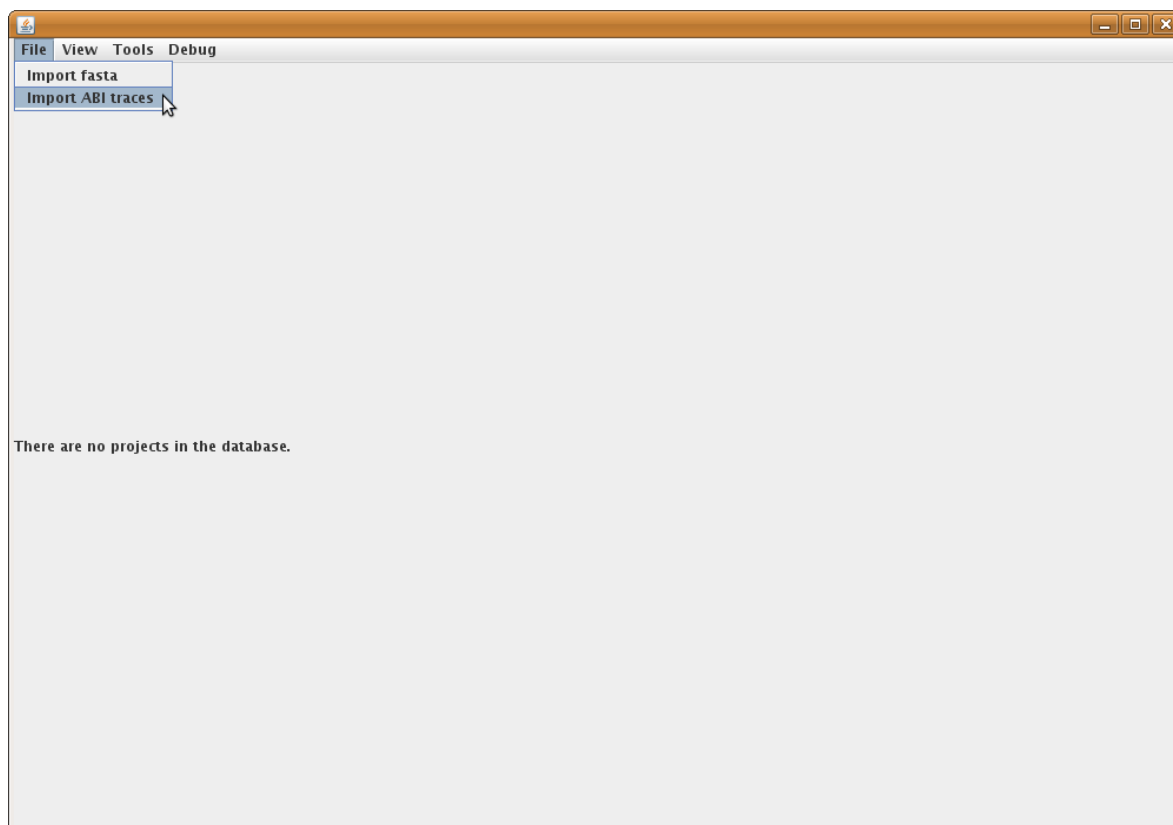
2.1.2 On Linux

Start the software by going to the application folder and typing: “java -jar ART-A.jar”.

3 Getting started

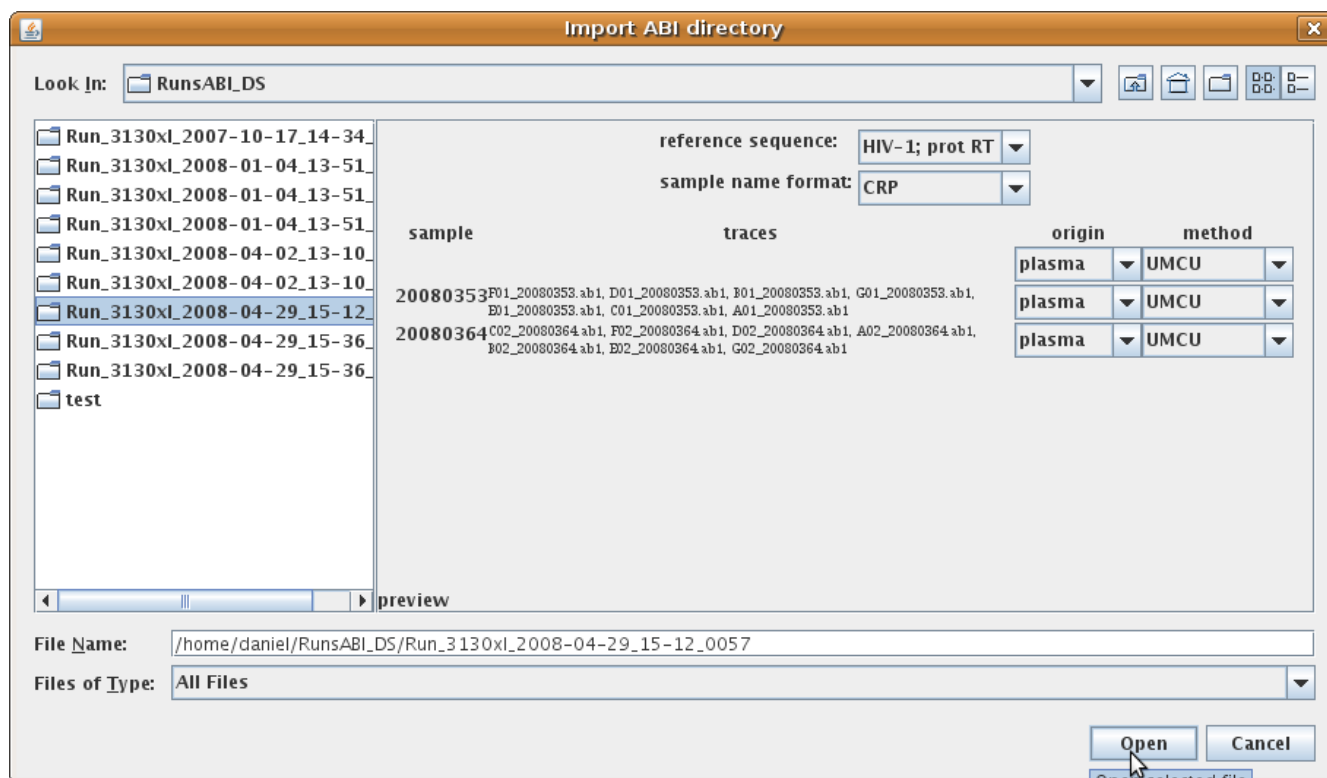
You have two possibilities to import sequences into the ART-A software:

- **Import ABI traces from a sequencer** [*File > Import ABI traces*]. For the follow-up go to page 6.
- **Import sequences from a fasta file** [*File > Import fasta*]. All the sequences from one fasta file will be saved in a project. For further details about the project view see page 13.



4 Importing ABI traces

Start the application and select [*File > Import ABI traces*]. The file menu to select the directory containing your ABI traces will open.



Two imported options must be set:

1. **reference sequence:** this is the part of the genome which has been sequenced
2. **sample name format:** used to retrieve the sample name from the filename.

(note: these options can be set permanently as default options in the preferences menu, see page 16)

Optional information about the sequenced samples:

1. origin: origin of the sample (DBS, DPS, blood, cell DNA, plasma, ...)
2. method: sequencing protocol used (UMCU, South Africa, Virco, ...)

Then select the appropriate directory containing the ab1 files you want to import from the left panel. You can select multiple directories by maintaining the control button (Ctrl).

While selecting the directory an overview of the detected samples will appear in the right panel. Proceed by clicking on the „Open“ button.

4.1 Base-calling and generation of the consensus sequence

At this stage the sequences are available as chromatograms only. Bases must be called to generate DNA sequences from the traces. In other words the peak heights of the chromatograms will be transformed into a sequence of nucleotides by an algorithm. This step is performed automatically when the ABI traces are imported.

The TraceTuner package ensures the base-calling and provides calibration tables for the following sequencer configurations:

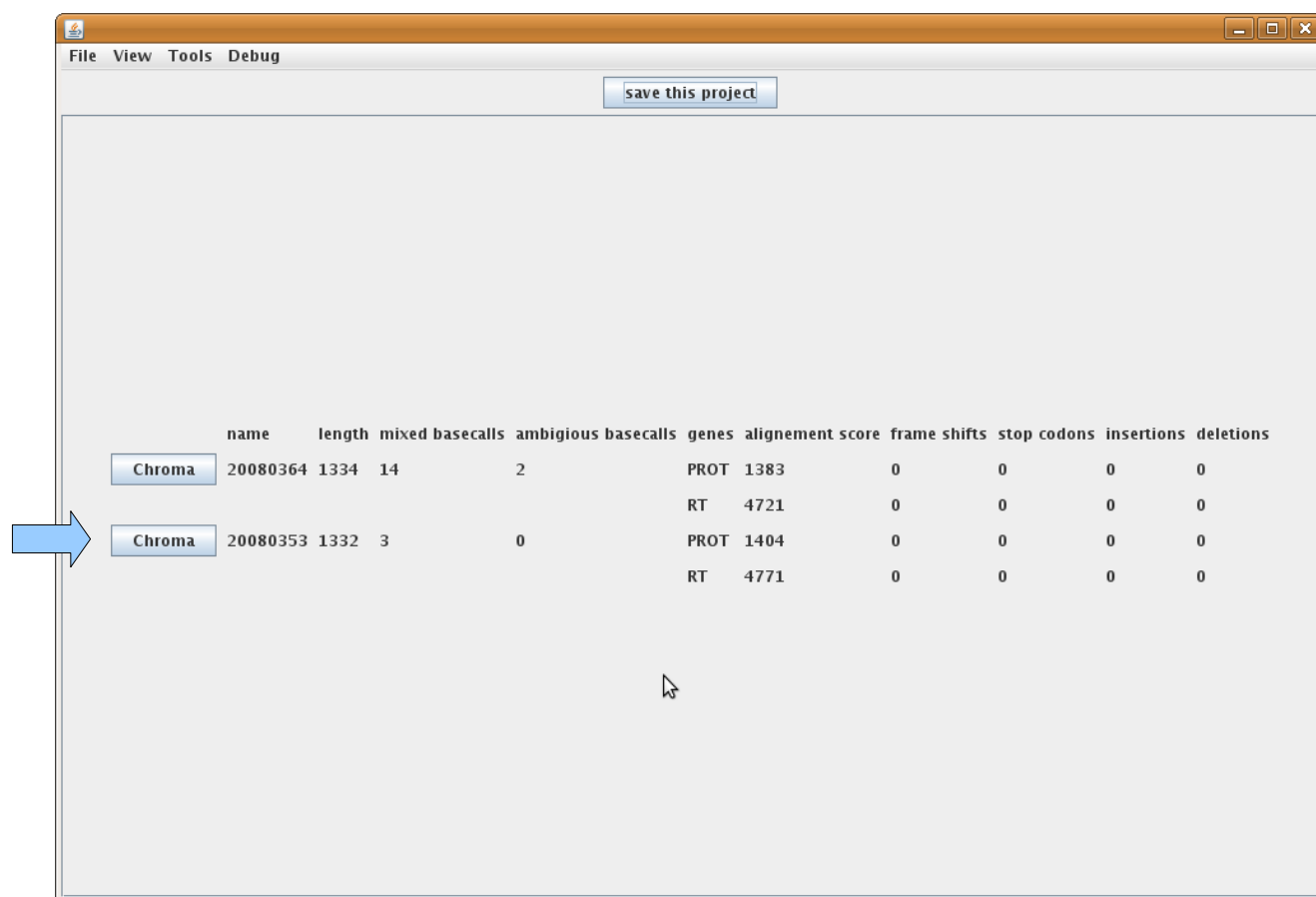
- 3100 (usable also for the 3130)
- 3730
- 3700 pop6
- 3700 pop5

TraceTuner also provides tools to generate custom lookup tables, provided enough sample data are available.

A consensus sequence is generated from the information of all the traces from a sample. This step is also performed automatically.

4.2 Base-call overview

After the consensus sequences have been generated for all the samples, an overview of the imported samples from the ab1 traces will be displayed.



The screenshot shows the TraceTuner software window with a menu bar (File, View, Tools, Debug) and a 'save this project' button. Below is a table with the following columns: name, length, mixed basecalls, ambiguous basecalls, genes, alignment score, frame shifts, stop codons, insertions, and deletions. The table contains two rows of data, each with a 'Chroma' label in a blue box to its left. A blue arrow points to the first 'Chroma' label.

	name	length	mixed basecalls	ambiguous basecalls	genes	alignment score	frame shifts	stop codons	insertions	deletions
Chroma	20080364	1334	14	2	PROT 1383	4721	0	0	0	0
Chroma	20080353	1332	3	0	PROT 1404	4771	0	0	0	0

If you are satisfied with the results, you can save them in a project. The project overview is explained on page 13.

4.3 Trace visualization

Sample traces can be visualized from the base-calling overview by pushing the “Chroma” button.



overview: returns to the overview

map: shows the position relative to the gene map of HIV-1

ambiguous basecalls: the cursor jumps to the next ambiguous basecall

mixed basecalls: the cursor jumps to the next mixed basecall

ref sequence: selected reference sequence

position: with respect to HXB2

amino acid: translation of the reference sequence, numbering starts at the beginning of the protein

20080364: sample name

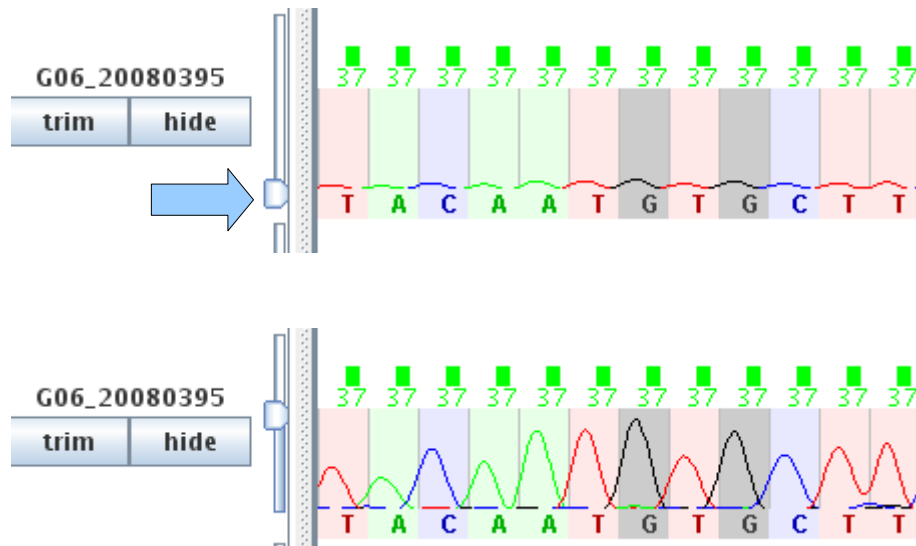
C02_20080364: name of the trace

trim: trim the trace

show/hide:	show or hide the chromatogram
consensus:	inferred consensus sequence
user editable:	modifiable consensus sequence
AA:	translation of the user editable sequence (will also show the influence of mixed basecalls by showing all possible amino acids)

4.3.1 Zoom slider

Adjust the zoom slider for optimal trace visualisation.



4.4 Trimming

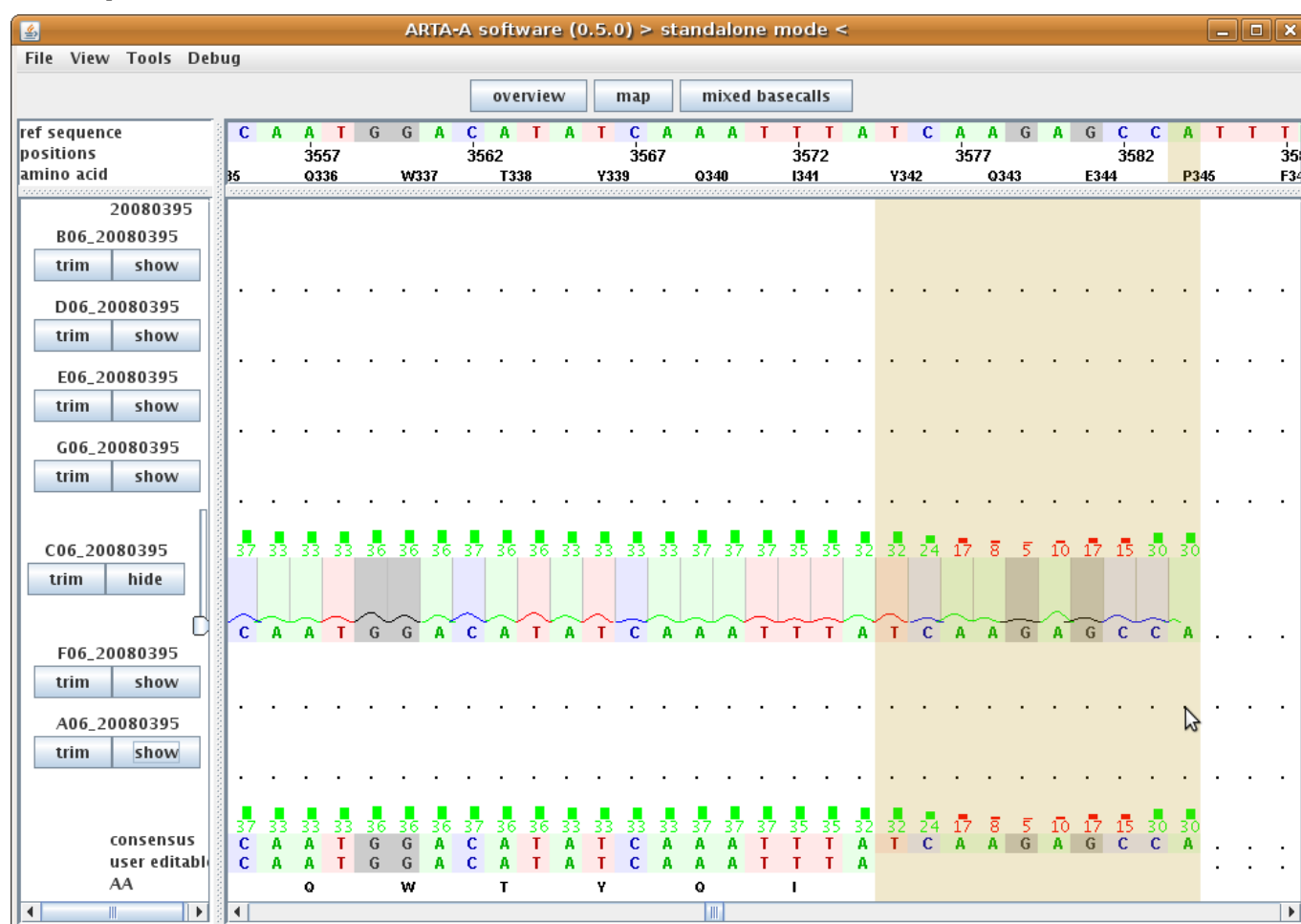
4.4.1 Automatic trimming

Automatic trimming is applied to the consensus sequence.

The default values are set to 10 for the trimming window and 20 for the trimming threshold: in other words, windows of 10 bases are analysed and the average quality values below 20 are discarded automatically. The trimming starts at the beginning and at the end of the consensus sequence.

The default preferences for the trimming can be set in the base-calling preferences, see page 16.

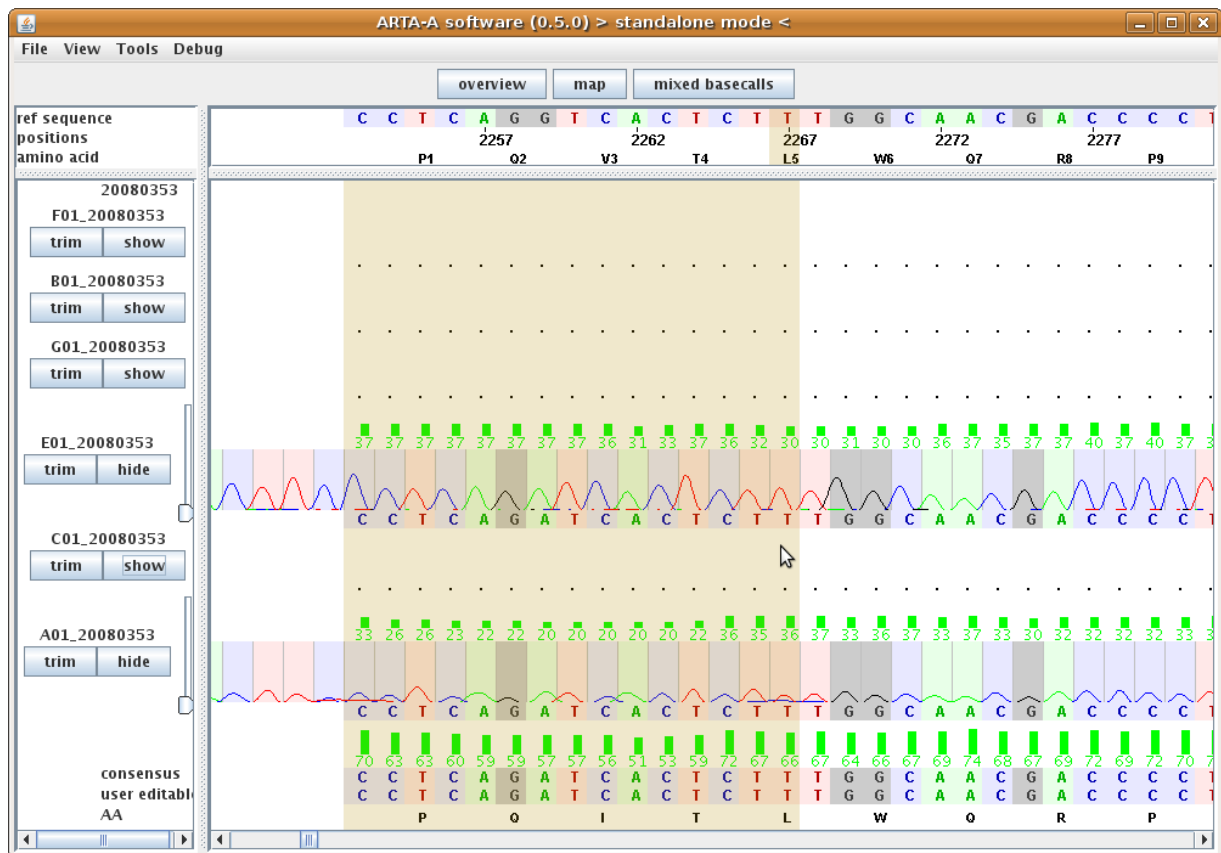
Example:



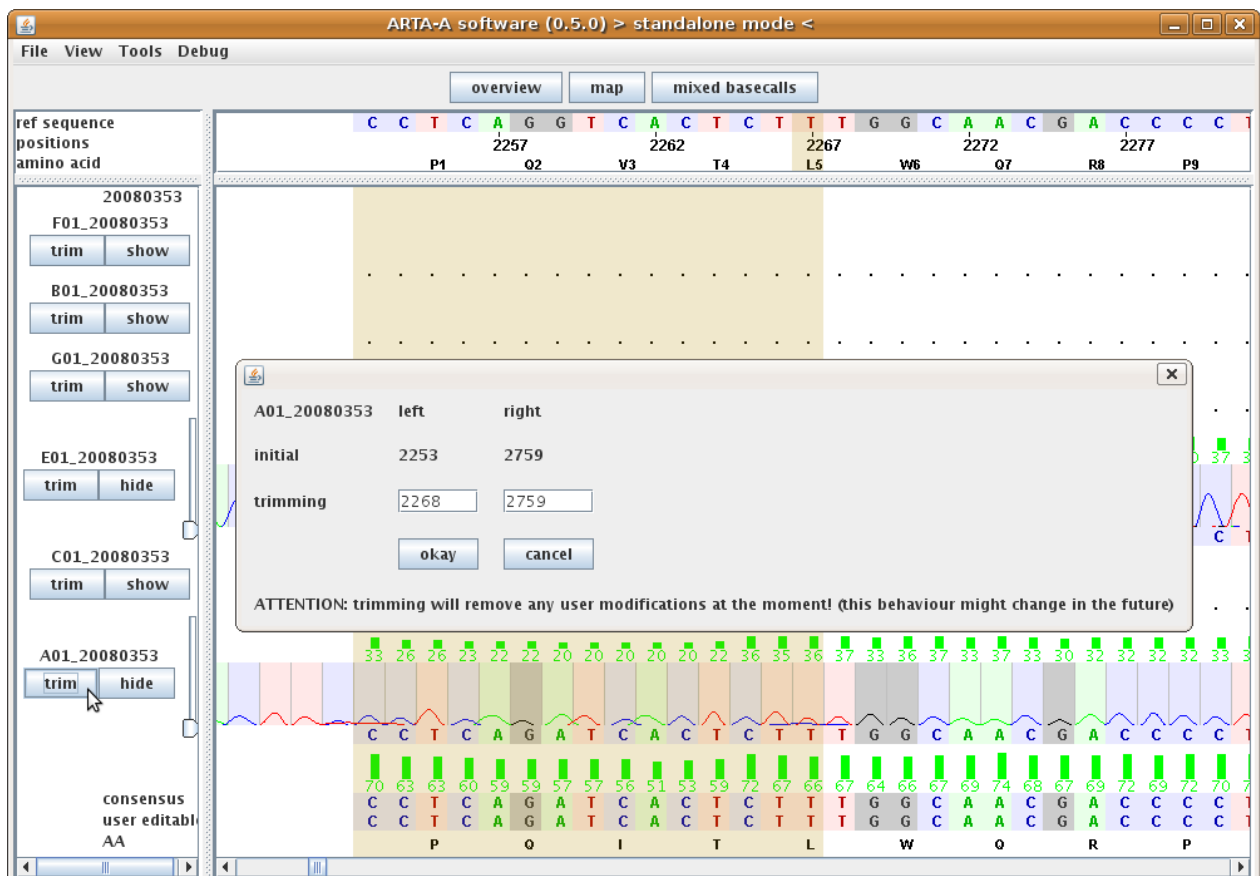
4.4.2 Manual trimming

You can manually trim the traces.

Situation before trimming:

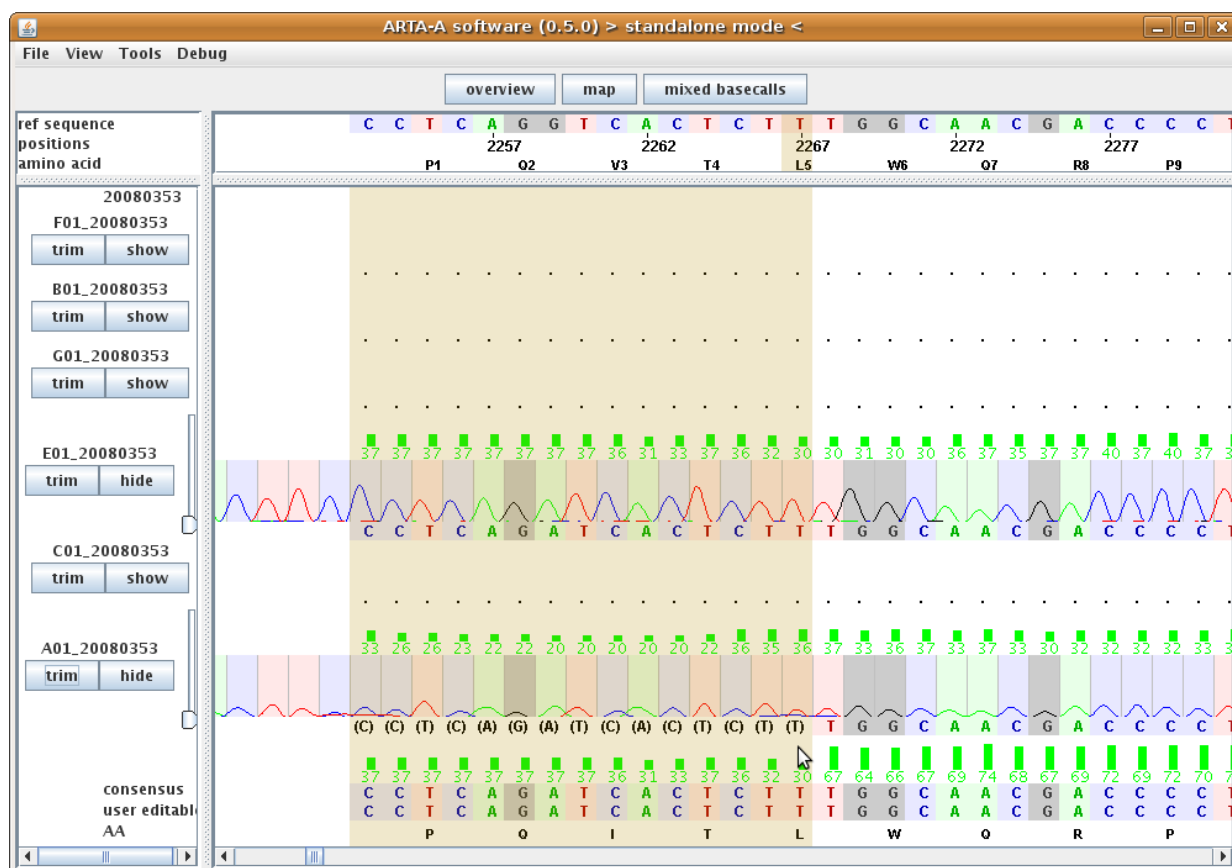


Select the region you want to omit from trace A01:



Result:

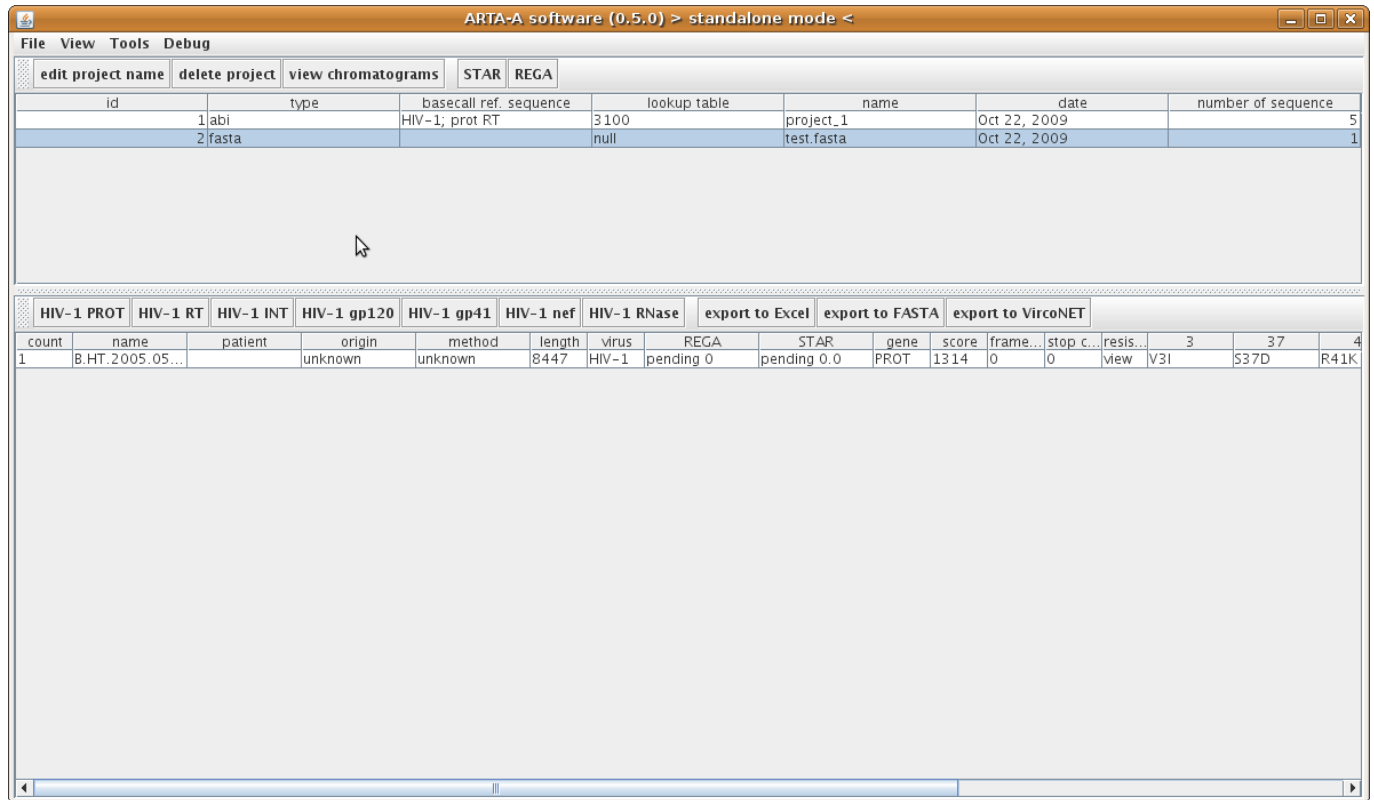
Remark the drop in the base-calling quality for the consensus sequence at the beginning as trace A01 is no more taken into account for region from 2253 to 2267.



5 Project view

All sequences stored in a project, whether they come from a sequencer or a fasta file, are subjected to the following analysis:

- Identification of the virus (HIV-1, HIV-2 and HCV at the moment).
- Identification of the genes (prot, RT, int, gp120, ...).
- Inference of mutations with respect to the reference sequence.



All projects stored in the database appear in the upper part of the window. The selected project is highlighted in light blue and the sequences contained in the project are displayed in more detail in the bottom panel.

Buttons of the upper panel:

- “edit project name” allows to change the project name
- “delete project” project will be removed completely from the database
- “view chromatograms” traces of the project can be (re)viewed.
- “STAR” initiate subtyping with the STAR subtyping tool (internet connection required)
- “REGA” initiate subtyping with the REGA subtyping tool (internet connection required)

Buttons of the bottom panel:

The first buttons show the detected genes of your sequences. Clicking on them will show the mutations associated with this gene.

“export to Excel” export the list of mutations to an Excel sheet

“export to FASTA” export the sequences in fasta format

“export to VircoNET” export the project in a format suitable for VircoNET tools

5.1 Subtyping

All the sequences of a project can be subtyped with the REGA

(<http://www.bioafrica.net/subtypetool/html/>) or STAR

(<http://www.vgb.ucl.ac.uk/subtyping.shtml>) subtyping tool. The sequences will be sent to the corresponding webserver and the results automatically retrieved and processed.

To initiate the subtyping press the “STAR” and/or “REGA” button and the process will be launched for the selected project.

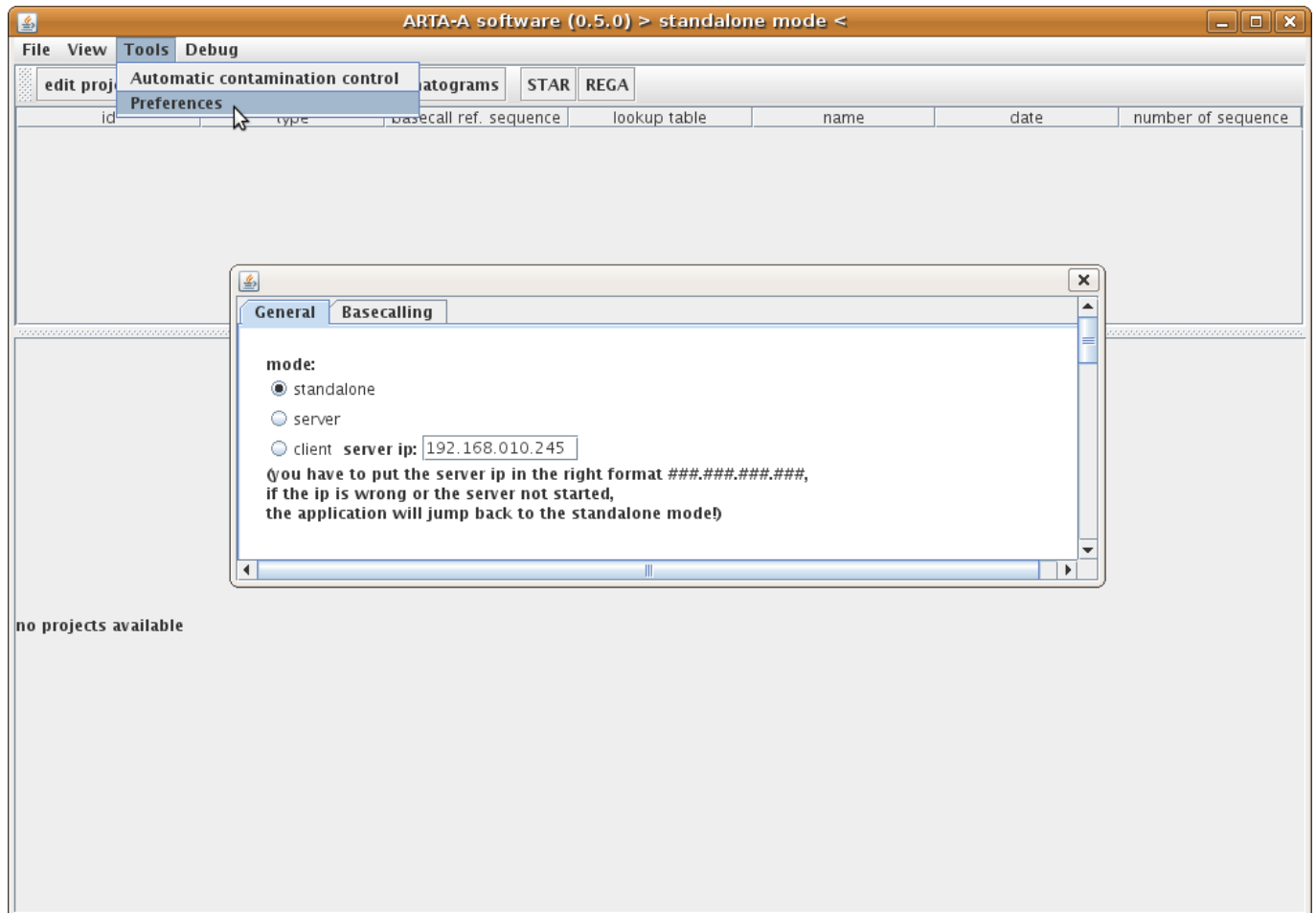
id	type	basecall ref. sequence	lookup table	name	date	number of sequence
2	abi	HIV-1; prot RT	3100	default	Oct 22, 2009	5

count	name	patient	origin	method	length	virus	REGA	STAR	gene	score	frame..
1	20080395		plasma	UMCU	1332	HIV-1	HIV-1 Subtype B 100	B 3.29034	PROT	1343	0
2	20080364		plasma	UMCU	1334	HIV-1	HIV-1 Subtype B 100	B 3.25502	PROT	1383	0
3	20080353		plasma	UMCU	1332	HIV-1	HIV-1 Subtype B 100	B 3.23328	PROT	1404	0
4	20080286		plasma	UMCU	1334	HIV-1	HIV-1 Subtype B 100	B 3.2314	PROT	1377	0
5	20080282		plasma	UMCU	1335	HIV-1	HIV-1 Subtype C 100	C 3.31735	PROT	1152	0

6 Preferences

Click on [Tools>Preferences] to open the preferences dialog.

6.1 General preferences



- standalone mode: default mode. the database, containing the projects, is stored in the folder of the application.
- server mode: clients can access and manipulate the projects saved on the server. The database is stored on the computer running in server mode.
- client mode: the application can connect to an installation running in server mode. The local database on the client is not used.

6.2 Base-calling preferences

Select the „Basecalling“ tab to view the base-calling preferences.

General **Basecalling**

sample name format:

☒ CRP A03_20030446_V3loop_KK1_1.ab1

☐ South Africa 2008-12-24_241208PASER_CP076302342_.B.ab1

☐ UMCU SAI_[sample_name].ab1

☐ Virco [sample_name].[plate ID]+[primer name].[...].ab1

default sample origin/type:

☐ unknown

☒ plasma

☐ blood

☐ DBS

☐ DPS

☐ cell DNA

default genotyping method:

☐ unknown

☒ UMCU

☐ South Africa

☐ Virco

default reference sequence:

☒ HIV-1; prot RT

☐ HIV-1; RT

☐ HIV-1; gp120

☐ HIV-1; gp41

☐ HIV-1; gp160

☐ HIV-1; nef

☐ HIV-1; RNase

trimming parameters:

trim window size: (default: 10, press enter to save the value)

trim threshold: (default: 20, press enter to save the value)

basecall lookup table:

☐ mbace built-in MegaBACE lookup table

☒ 3100 Use the built-in ABI 3100-pop6 lookup table

☐ 3730 built-in ABI 3730-pop7 lookup table

☐ 3700pop6 ABI 3700-pop6 lookup table

☐ 3700pop5 ABI 3700-pop5 lookup table

☐ custom custom made lookup table

sample name format	used to retrieve the sample name from the filename
sample origin	origin of the sample
genotyping method	sequencing protocol used
reference sequence	part of the genome or gene(s) which has been sequenced
trimming parameters	see page 10
basecall lookup table	define sequencer and chemicals used

6.3 Contamination control preferences

Not yet implemented.

7 Advanced tools

7.1 Contamination control

[Tools > Automatic contamination control]

This algorithm tries to find mislabelled samples by comparing every sample of a patient with every other sample in the database.

The following steps are performed:

1. quick search based on similarity to find possible candidates.
2. for every candidate, a multiple alignment is performed with the reference sequence from the Los Alamos database, every sequence of the candidate and sequences close to the candidate from the ART-A software database.
3. Phylogenetic trees are generated from the alignments.
4. The patients sequences are tested for monophyly on the tree. If not monophyletic, a sample could have been mislabelled for this patient.

Inspect the results by analysing the phylogenetic trees visualised by FigTree.

7.2 Access to data

All the data of the projects are stored in a database located in the folder "data/database".

With [*Debug>H2 Webserver*] you can access your data through a web interface. The data can be queried with the SQL language.

8 Definitions

ambiguous basecall	the base-calling algorithm cannot decide between two bases.
calibration table	used by the base-calling algorithm to adapt to different sequencer configurations.
chromatogram	visual output of the chromatograph.
DBS	d ried b lood s pot
DPS	d ried p lasma s pot
mixed basecall	due to the population sequencing technique, two or more bases can be possible at one position.
trimming	parts of a chromatogram can be of poor quality and need to be excluded from the calculation of the consensus sequences. This is done automatically but can also be manually set.

8.1 Acknowledgements

I would like to thank the following persons for their contribution to the ART-A software project:

Danielle Perez-Bercoff, researcher at CRP-SANTE in Luxembourg, for guidance in the project.

Gennady Denisov, the author of TraceTuner, for helping to adapt TraceTuner to the ART-A software.

Anne-Marie Ternes, bioinformatician at CRP-SANTE in Luxembourg, for testing the application and reviewing this document.

Carole Wallis, Sue Aitken and Michelle Bronze for testing the application.

My colleges at the Retrovirology Laboratory of Luxembourg for testing the application and particularly Christine Lambert and Jean-Yves Servais.