ATTILA User Guide

University of Brasilia
2016

About ATTILA

ATTILA (AutomaTed Tool For Immunoglobulin Analysis) searches for candidate immunoglobulin sequences in phage display libraries, generating as main output a list of sequences of heavy and light chain, which were selected by phage display experiment, and code for antibody fragments that can probably bind to the target molecule. ATTILA package has programs developed in C, Perl and Shell script to execute eight steps of a completely automated analysis (Figure 1). The third-party tools used by ATTILA are listed in section **Requirements**.

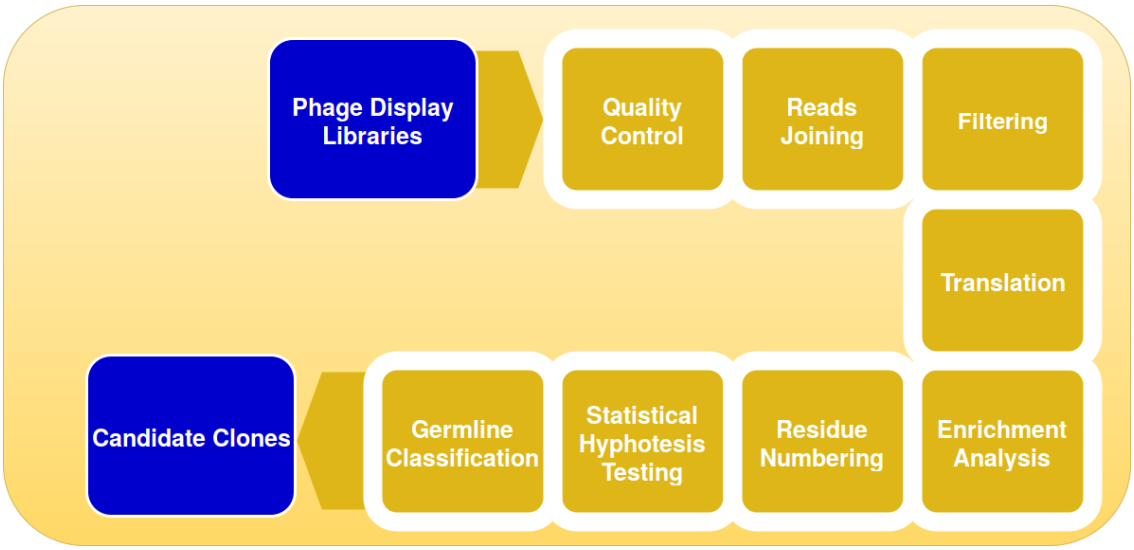

Figure 1: Analysis performed by ATTILA.

ATTILA can analyse human sequences coding for variable domain of heavy (VH) and light chain (VL), produced by phage display technology. Therefore, the input for the method must be VH and VL libraries, from initial and final rounds. Considering that our approach uses distances between canonical aminoacid residues of variable domain based on human sequences, the analysis may also be performed on mouse sequences, since their distances are similar to those of human.

The package has a very simple structure, with two directories, called **data** and **programs**, together with this user guide and an example of configuration file. The directory **data** contains human databases from NCBI, used by IgBlast to perform germline classification, while **programs** directory keeps all scripts and programs of the method.

Requirements

In order to install and run ATTILA, you must have:

- Linux system
- FASTQC `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`
- Prinseq-lite `http://prinseq.sourceforge.net/`
- FastqJoin `https://code.google.com/archive/p/ea-utils/`
- Perl `https://www.perl.org/get.html`
- IgBlast `ftp://ftp.ncbi.nih.gov/blast/executables/igblast/release/`
- R package `https://cran.r-project.org/`
- ggplot2 Use R function install.packages("ggplot2")
- scales Use R function install.packages("scales")
- Internet

Installation

After installing all requirements, perform the following steps:

1. Download ATTILA package at ...
2. Uncompress the `tar.gz` file using command line:

   ```
   tar -vzxf attila-1.0.tar.gz
   ```

3. Go to the directory where you want to install ATTILA, using `cd` command. Note that ATTILA and IgBlast packages must be subdirectories of the installation directory.
4. Type the following command line:

   ```
    ln -s <path to check_requirements.sh> check_requirements.sh
   Example: ln -s home/Attila/programs/check_requirements.sh check_requirements.sh
   ```

   Remember that check_requirements.sh is located in a directory called "programs" of ATTILA package.

5. Run the following command line:

   ```
   ./check_requirements.sh
   ```

If check_requirements.sh prints "Type ./attilacli.sh to run ATTILA", then ATTILA is ready to run ! If not, check_requirements.sh prints a list of requirements that still need to be installed.

Getting started

Starting ATTILA

ATTILA uses two configuration files to run the analysis, one for the VH library and one for the VL library. These files may be manually created or you may let ATTILA do it for you. In the first case, you can use our example files called **SingleEndReads_VH.cfg** and **SingleEndReads_VL.cfg** (for single-end reads) or **PairedEndReads_VH.cfg** and **PairedEndReads_VL.cfg** (for paired-end reads) located in the parent directory of ATTILA package. Just copy the files and change configurations according to your data. Note that the parameter called "Project Name" must be the same for both configuration files. The other way is to create the configuration files is to answer some questions asked by ATTILA. It will automatically generate the configuration files at the end of the process.

To start ATTILA, type the command line:

```
./attilacli.sh
```

Running ATTILA when you already have the configuration files

ATTILA will ask if you already have the configuration files. If you have, type "y". ATTILA will ask the settings file path for VH and for VL. **Note that the analysis is executed separately for VH and VL, therefore, you must have two configuration files, one for each library type.**

Using ATTILA to create configuration files

If you prefer to let ATTILA create the configuration files, type "n" or press the ENTER key when ATTILA asks if configuration files already exist. ATTILA will ask you some questions in order to fill the parameters for the VH and VL libraries. This step may be time consuming for first-time users but, at the end, ATTILA will execute the complete analysis of VH and VL libraries for you. Here is the complete list of parameters asked by ATTILA:

- Project name: Name of the directory that will created by ATTILA to save output files
- Directory to save the project: The directory where the project will be saved
- Reads are paired-end: If yes, type "y" or press ENTER key. If not, type "n"

If your reads are paired-end, ATTILA will ask the location of all eight input files, using the following parameters:

- VH R0 reads r1 path: location of the `fastq` file containing reads r1 from initial VH library
- VH R0 reads r2 path: location of the `fastq` file containing reads r2 from initial VH library

- VH RN reads r1 path: location of the `fastq` file containing reads r1 from final VH library
- VH RN reads r2 path: location of the `fastq` file containing reads r2 from final VH library
- VL R0 reads r1 path: location of the `fastq` file containing reads r1 from initial VL library
- VL R0 reads r2 path: location of the `fastq` file containing reads r2 from initial VL library
- VL RN reads r1 path: location of the `fastq` file containing reads r1 from final VL library
- VL RN reads r2 path: location of the `fastq` file containing reads r2 from final VL library

Initial library is the library sequenced before phage display experiment. Final library is the library sequenced after all rounds of phage display.

If your reads are single-end, ATTILA will ask the location of four input files, using the following parameters:

- VH R0 path: location of `fastq` file containing reads from initial VH library
- VH RN path: location of `fastq` file containing reads from final VH library
- VL R0 path: location of `fastq` file containing reads from initial VL library
- VL RN path: location of `fastq` file containing reads from final VL library

The remaining parameters will be asked for both types of reads:

- Minimum read length: the default value is 300 pb (approximate size of variable domain coding region); type "y" to change default value and enter the new read length using an integer number ; if want to use the default, type "n" or press ENTER key
- Minimum base quality: the default value is 20; type "y" to change default value and enter the new base quality using an integer number ; if want to use the default, type "n" or press ENTER key
- Number of candidates to rank: number of candidate clones that ATTILA will try to find in VH and VL libraries; the number must be an integer

ATTILA will print all configurations you have entered, so that you can check if they correct. In positive case, type "y" or press ENTER key. Then, ATTILA will name the configuration files using the project name and "VH" or "VL" ending, eg.: myproject_VH.cfg and myproject_VL.cfg. If you need to correct anything, type "n" and the configuration editing menu will be open. You can learn how to use this menu in section **Configuration editing menu**.

Running the analysis

After you have entered the configuration files path or the parameters necessary to create them, ATTILA will start the analysis of your data. It will print some messages to inform what is currently being done, when the analysis of both libraries are completed and the execution time of the analysis. In case of a successful analysis, the following messages will be printed:

```
Creating project directory
Running VH analysis ...

real 1m21.031s
user 1m24.390s
sys  0m3.320s
--------------------------------
VH Analysis Completed
--------------------------------
Running VL analysis ...

real 2m42.973s
user 2m42.060s
sys  0m6.410s

--------------------------------
VL Analysis Completed
--------------------------------
Analysis report is ready !
```

Visualizing results

ATTILA creates in your project directory separate directories for VH and VL libraries[1], where most of the files generated will be located, as shown in Figure 2. A summary of the analysis of both libraries can be found in Report directory, which is also a subdirectory of the project directory.

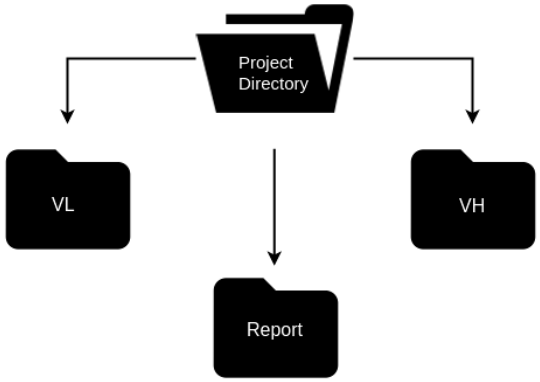

Figure 2: Project directory structure.

For each library type, ATTILA creates the same subdirectory structure (Figure 3), with 3 subdirectories: InitialRound, FinalRound and SelectedSequences. The files located in the InitialRound directory are intermediate output files generated by the analysis of the initial library, the library sequenced before phage display experiment. In the FinalRound directory, the files are from the final library, the library sequenced after the experiment. And the SelectedSequences directory has the files containing the results of the analysis, i.e., information about the candidate clones of the corresponding library type (VH or VL).

What files really matter?

Even though a lot of files are generated by the analysis, you do not need to see all of them. It is worthy to mention that in the file name, the "?" character is the number of candidate clones chosen by you. So here are the files that really matter to you:

- **Report.html** The analysis report for VH and VL, showing germline classification, fold change, statistics, regions of variable domain of candidate clones and reads information. This file is located in **Report** directory. You will need Internet connection to correctly visualize this report, since its content is dinamic[2]
- **vhlist?numbered.fasta** A `fasta` file, located in **SelectedSequences** subdirectory of VH, containing the aminoacid sequences from VH candidate clones.

[1]In the project directory there are also `log` files, with standard output from the programs used in the method. These files are just a way for us to keep track of possible errors, so you do not need to check on them.

[2]This report of the analysis will be explained in detail in section **Better explained analysis report.**

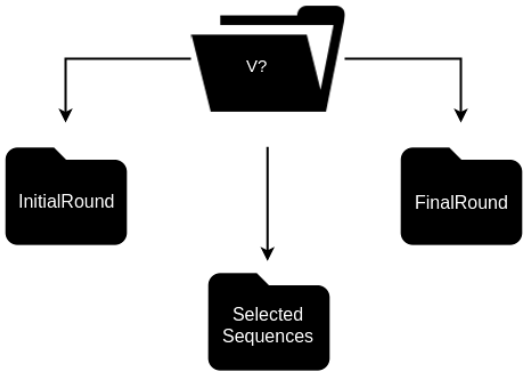

Figure 3: Library directory structure. The "?" can be a "H" (heavy chain) or a "L" (light chain).

- **vllist?numbered.fasta** A `fasta` file, located in **SelectedSequences** subdirectory of VL, containing the aminoacid sequences from VL candidate clones.
- **vhlist?numberednt.fasta** A `fasta` file, located in **SelectedSequences** subdirectory of VH, containing the nucleotide sequences from VH candidate clones.
- **vllist?numberednt.fasta** A `fasta` file, located in **SelectedSequences** subdirectory of VL, containing the nucleotide sequences from VL candidate clones.
- **vhSequenceCounting.csv** A csv file, located in **VH** directory, containing the number of reads of each step of the analysis.
- **vlSequenceCounting.csv** A csv file, located in **VL** directory, containing the number of reads of each step of the analysis.
- ***fastqc.zip**: reads quality reports generated by FASTQC, they are in InitialRound and FinalRound directories, for both library type directories, VH and VL. Note that the evaluation of reads quality is done with initial and final libraries before and after filtering step, so for each library there are two FASTQC reports.

Better explained analysis report

The analysis report gather different results in one file, allowing a friendly visualization of your data. The report has 2 tabs, one for each library type (VH and VL). In each tab, there are 3 sections: **Reads Information**, **Candidate Clones** and **Regions of Variable Domain of Candidate Clones**.

The section **Reads Information** has one table showing the loss of reads of initial and final libraries (Figure 4). The loss of reads is the percentage of reads removed after filtering reads by length and base quality score PHRED. All reads with less than 300 pb and/or less than 20 base quality score are removed. If you have set ATTILA with different values of

minimum length and base quality, then the filtering step will take these values as thresholds to remove reads.

Besides the loss of reads table, the section has two plots (Figure 4). One plot shows the proportion of reads with adequate and inadequate length, i.e., the proportion of reads with more than or equal to the minimum read length (default is 300 pb) and the proportion of reads with less than the minimum read length. The nice thing about this plot is the rationale behind it. Considering that the filtering step removes reads with inadequate length and/or inadequate base quality, if you had a high loss of reads and the FASTQC report shows that the reads have good sequencing quality, then your reads were removed because of inadequate length. In this case, you will see in the plot a proportion of reads with inadequate length bigger than that of reads with adequate length. In summary, this plot gives you a hint of what happened in the filtering step.

The other plot shows the number of reads by task, i.e., the number of reads in each step of the analysis. The number of reads of task called "None" are in fact the raw data, the reads before any processing. The number of reads is shown for both, initial and final library. The only task that does not have a quantity for both libraries is the "Enrichment", since this step compares initial and final library to find clones with increased frequency, and produces as output a file with sequences from final library to represent the enriched clones.

Figure 4: Section Reads Information of the analysis report.

The section **Candidate Clones** has one table showing information of the candidate clones (Figure 5), such as the fold change (how much the frequency increased from initial to final library), the p-value (if it is less than α, then the difference of proportion of a given clone is statistically significant), confidence interval (the narrrower the interval the more precise are the clone proportions we calculated), the germline that the candidate clone belongs to

and its respective alignment identity value. If you hover the mouse on the rank number, you will see the id of the candidate sequence. It is worthy to mention that the number of candidate clones found may be less than the number of candidates you asked, because it is possible that not all the most frequent clones have the canonical aminoacid residues of immunoglobulin variable domain, and therefore they can not be considered candidates according to our biological criterion.

Candidate clones

Sequence ID	Fold Change	p-value (αBonf=0.005)	InfIC (IC=95%)	SupIC (IC=95%)	Germline	Identity (%)
1°	33335.35	0	-0.6905875	-0.6813375	VH1-8	98.913
2°	129.37	0	-0.003156721	-0.002126634	VH1-8	100
3°	82.90	0	-0.002098522	-0.001272144	VH1-8	100
4°	40.19	5.686198e-10	-0.001095816	-0.0005172448	VH1-8	98.99
5°	33.91	1.421599e-08	-0.0009435148	-0.0004110743	VH1-8	98.99
6°	30.14	9.867993e-08	-0.0008511173	-0.0003483889	VH1-8	98.99
7°	22.61	4.848389e-06	-0.0006633065	-0.000226034	VH1-8	98.99
8°	20.10	1.789701e-05	-0.0005995513	-0.0001864006	VH1-8	98.99
9°	20.10	1.789701e-05	-0.0005995513	-0.0001864006	VH1-8	97.826
10°	15.07	0.0002480892	-0.0004696288	-0.000109546	VH1-8	100

Regions of Variable Domain of Candidate Clones

Figure 5: Section Candidate Clones of the analysis report.

Finally, the section **Regions of Variable Domain of Candidate Clones**, has one table showing the aminoacid residues from each region of variable domain of the candidate clones (Figure 6). The sequences ids are also shown if you hover the rank number. The sequences used by IgBlast to identify each region of variable domain of candidate clones are germline sequences, downloaded from IgBlast webpage [1, 2].

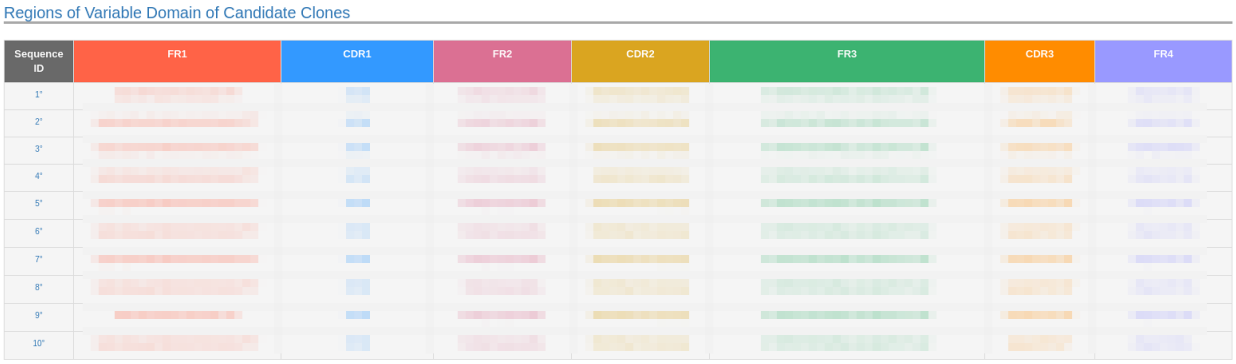

Figure 6: Section Regions of Variable Domain of Candidate Clones of the analysis report.

Configuration editing menu

If you let ATTILA create the configuration files, after you have entered all parameters, it will print the configuration and ask if it is correct. Type "y" or press ENTER key, if the configuration is correct. Then ATTILA will start the analysis. If the configuration is incorrect, type "n" to open the configuration editing menu. In this case, the following list of parameters will be printed:

```
--------------------Configuration Editing Menu--------------------------

Project Name (1)
Directory to save project (2)
Reads are paired-end (4)
Path of fastq file of VH R0 paired-end reads r1 (5)
Path of fastq file of VH R0 paired-end reads r2 (6)
Path of fastq file of VH RN paired-end reads r1 (7)
Path of fastq file of VH RN paired-end reads r2 (8)
Path of fastq file of VL R0 paired-end reads r1 (9)
Path of fastq file of VL R0 paired-end reads r2 (10)
Path of fastq file of VL RN paired-end reads r1 (11)
Path of fastq file of VL RN paired-end reads r2 (12)
Path of fastq file of VH R0 single-end reads (13)
Path of fastq file of VH RN single-end reads (14)
Path of fastq file of VL R0 single-end reads (15)
Path of fastq file of VL RN single-end reads (16)
Minimum Read Length (18)
Mininum Base Quality (19)
Number of Candidates (20)
Save and exit (0)
------------------------------------------------------------------------
Enter corresponding integer to correct settings:
```

You must enter the number corresponding to the parameter you want to change. Then ATTILA will ask you the new configuration. The parameter corresponding to number "4" is a bit different. If you type "4", ATTILA will tell you to type "1" if your reads are paired-end or "0" if your reads are single-end. When you are done editing the configuration, type "0" to save and quit the menu. ATTILA will create the configuration files and start the analysis.

Advanced Topic

What if you want to run ATTILA from a different directory?

As explained in section **Installation**, you must choose a directory to install ATTILA where the IgBlast and ATTILA packages are subdirectories. But, after you have installed it you may run ATTILA from another directory by doing the following instructions:

- Go to the new directory where you want to run ATTILA, using cd command.
- Type the command line:

  ```
  ln -s <path to attilacli.sh> attilacli.sh
  Example: ln -s /home/Attila/programs/attilacli.sh attilacli.sh
  ```

- Copy the file "paths_attila.txt" to the directory where you want to run ATTILA, using the command line:

  ```
  cp <path to file paths_attila.txt> .
  Example: cp /home/paths_attila.txt .
  Note that the dot (.) in the end of the command is necessary.
  ```

- Run attila typing:

  ```
  ./attilacli.sh
  ```

References

[1] NCBI. Igblast tool. `http://www.ncbi.nlm.nih.gov/igblast/`.

[2] Jian Ye, Ning Ma, Thomas L Madden, and James M Ostell. Igblast: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research*, page gkt382, 2013.