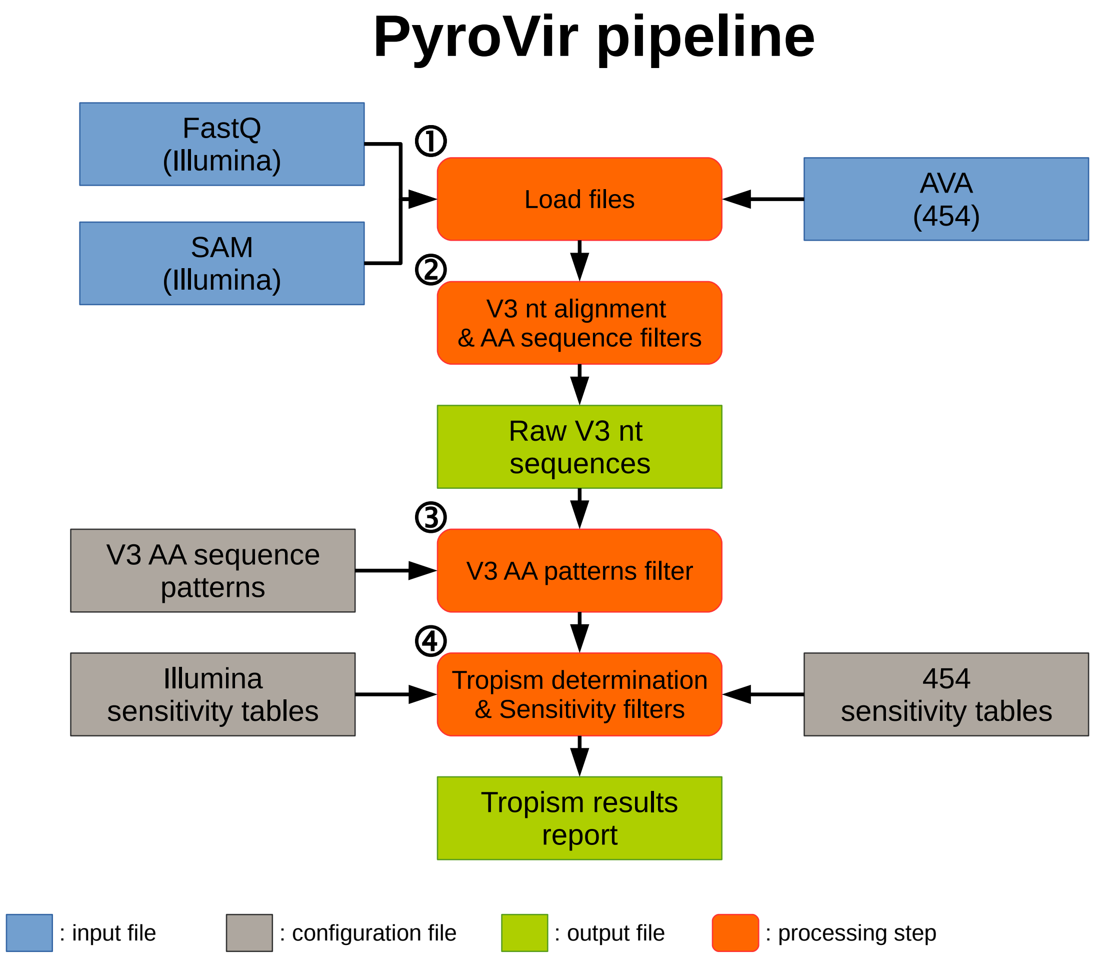
Purpose

The PyroVir software is an effective tool to interpret HIV Tropism next generation sequencing (NGS) information and guide the adaptation of anti-HIV treatment. The purpose of the software is to identify the tropism of an individual patient's HIV strains and shows whether the patient is infected with virus that enters cells using the CCR5 co-receptor, the CXCR4 co-receptor, or both.

Software pipeline



Step 1 :
The software is designed to process 454 and Illumina data files. For 454 technology the input files are AVA consensus reads output files (must contain galign_con_ prefix) and for Illumina technology FastQ or SAM files. If the subtype is known for the samples, this information can be added for a better tropism determination.

Step 2 :
The input sequences are aligned versus the BaL V3 sequence and trimmed at the boundaries of V3. The trimmed sequences are momentarily translated to amino-acid sequences and filtered. Those with no Cysteine as first amino-acid required for the disulfide bound maintaining the V3 loop, with a STOP codon, outside the size limits (between 33 and 37 AA by default) or containing ambiguous characters (N) in the nucleotide sequence are discarded.
The nucleotide remaining sequences are written in a fasta file. The sequences headers contain the count information :

```
>CON_2 Count=874
TGTACAAGACCCAATAACAATACAAGAAAAGGTATAGCTATAGGACCAGGGAGAACATTTTATACAAGAGAAAAAATAATAGGAAATATAAGACAAGCACATTGT
>CON_5 Count=758
TGTACCAGACCCAATAACAATACAAGAAAAAGTATAACTATAGGACCAGGGAGAGCATTTTATACAAGAGGAGAAATAATAGGAAATATAAGACTAGCACATTGT
>CON_3 Count=539
TGTACAAGACCCAATAACAATACAAGGAAAAGTATACCCATAGGACCAGGGAGAGCATTTTATACAAGAGGAGAAATAATAGGAAATATAAGACTAGCACATTGT
>CON_16 Count=270
TGTACAAGACCCAATAACAATACAAGAAAAAGTATACATATAGGACCAGGGGGAGCATTTTATACAACAGGAGGAATAATAGGAGATATAAGACAAGCACATTGT
>CON_8 Count=212
TGTACAAGACCCAATAACAATACAAAAAAAGGTATAGCTATAGGACCAGGGAGAACATTTTATACAAGAGAAAAAATAATAGGAAATATAAGACAAGCACATTGT
>CON_1 Count=202
TGTACAAGACCCAATAACAATACAAGAAAAAGTATACCCATAGGACCAGGGAGAGCATTTTATACAAGAGGAGAAATAATAGGAAATATAAGACTAGCACATTGT
>CON_91 Count=188
TGTACAAGACCCAATAACAATACAAGCAAAAGTATACCTATAGGACCAAGGAGAGCGTTTTATGCAACAGGAAGAATAATAGGAGATATAAGACAAGCATATTGT
...
```

Step 3 :

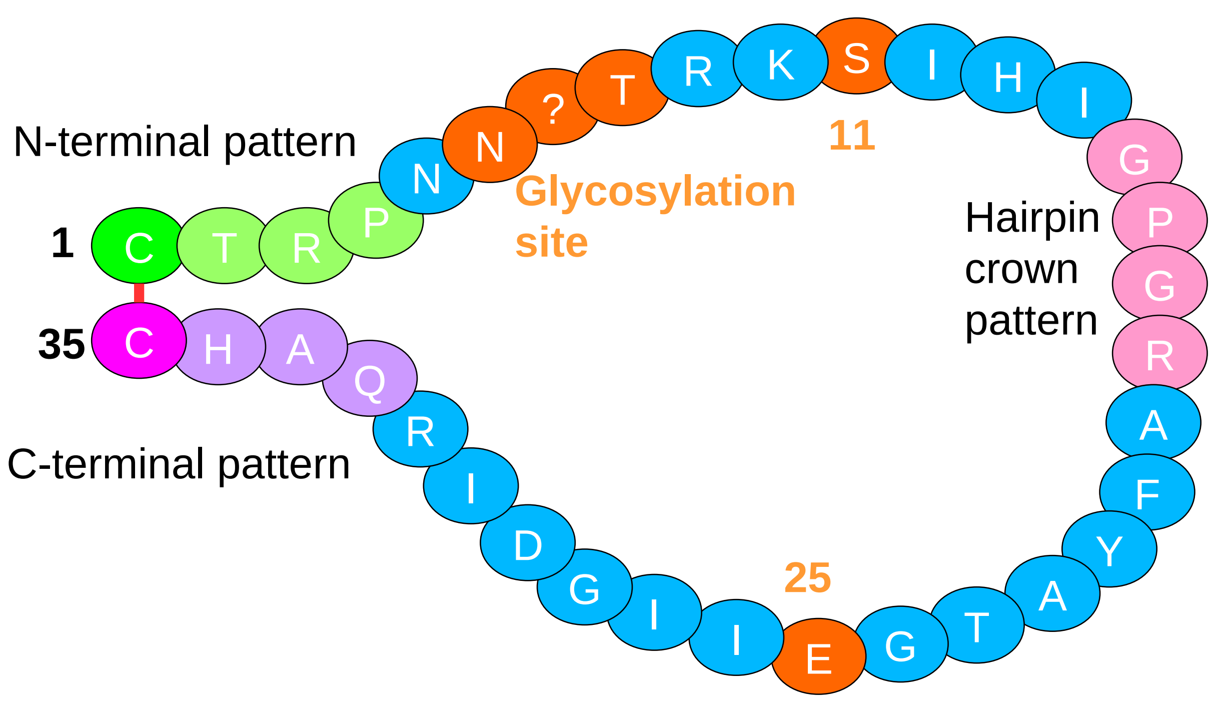
With the new fasta file, each sequence is translated and filtered on the bases of three V3 specific conserved patterns. The N-terminal pattern (CTRP), the hairpin crown pattern (GPGR) and the C-terminal pattern (QAHC). An identity of 0.5 and 0.75 was allowed for substitution and insertion/deletion of an amino-acid at any of the three signature patterns. Sequences which do not match with those patterns constraints are discarded.

Step 4 :

For each remaining amino-acid sequence, the tropism is determined based on the combined rule adapted to the subtype of the sample if this field was entered in step 1, if not, the B subtype combined rule is applied.

The combined rule is based on V3 specific positions amino-acids composition (as 11, 25, glycosylation sites…) and the net electrostatic charge of the sequence as shown on the V3 loop key positions figure.

V3 loop : Key positions



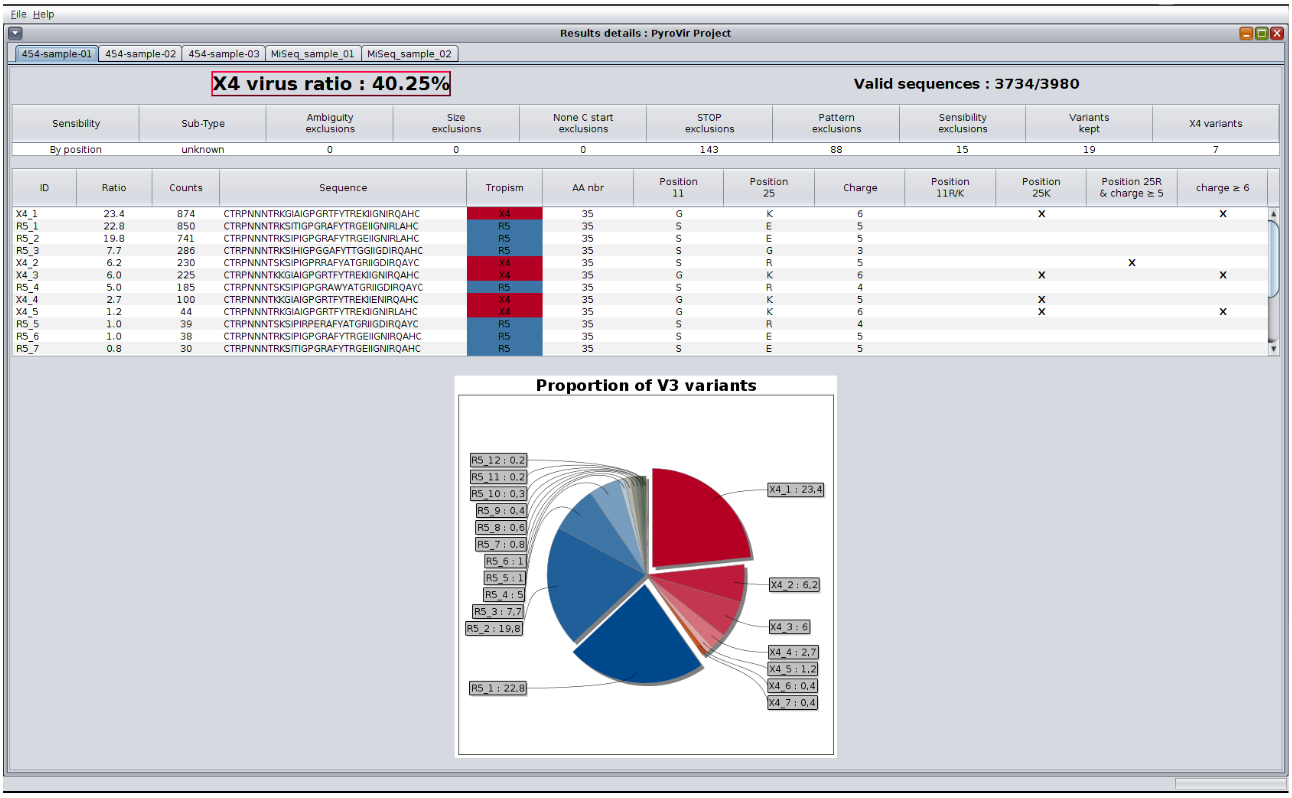
Net electrostatic charge = (∑R + ∑K) - (∑D + ∑E)

A stastistic filter is applied to each sequence to discard artifactual sequences resulting from PCR or

sequencing errors. Cut-off threshold are based on the results from 20 clones sequencing (Illumina and 454 technologies) which generate two matrices for each technology based on Poisson distributions. The first one is a V3 Global sensitivity with one column for the frequency values of the whole V3 sequence and the rows are the coverage. The second one is a Position-specific sensitivity matrix with all V3 positions as columns and coverage as rows.
The choosen Global and Position-specific tables for the studied sample are specific to the sequencing technology.

For R5 tropic sequences, the software uses the Global sensitivity table to discard the sequences which frequency is below the table cut-off depending on the coverage of the sequence. For X4 tropic sequences, the Position-specific table is used. The V3 position of the sequence that switch the tropism from R5 to X4 is the key position that is looked for in the column table and the sequence is discared if its frequency is below the cut-off depending on the coverage of the sequence.

Finally, a report is produced indicating the tropism X4 ratio and detailing the quasi-species with a pie-chart and a sequences table for each sample. The criteria determining the X4 tropism are presented in the sequences table. Also, the number of discarded sequences for each filtering steps are notified in the filter table.



X4 virus ratio : 40.25%

Valid sequences : 3734/3980

Sensibility	Sub-Type	Ambiguity exclusions	Size exclusions	None C start exclusions	STOP exclusions	Pattern exclusions	Sensibility exclusions	Variants kept	X4 variants
By position	unknown	0	0	0	143	88	15	19	7

ID	Ratio	Counts	Sequence	Tropism	AA nbr	Position 11	Position 25	Charge	Position 11R/K	Position 25K	Position 25R & charge ≥ 5	charge ≥ 6
X4_1	23.4	874	CTRPNNNTRKGIAIGPGRTFYTREKIIGNIRQAHC	X4	35	G	K	6		X		X
R5_1	22.8	850	CTRPNNNTRKSITIGPGRAFYTRGEIIGNIRLAHC	R5	35	S	E	5				
R5_2	19.8	741	CTRPNNNTRKSIPIGPGRAFYTRGEIIGNIRLAHC	R5	35	S	E	5				
R5_3	7.7	286	CTRPNNNTRKSIHIGPGGAFYTTGGIIGDIRQAHC	R5	35	S	G	3				
X4_2	6.2	230	CTRPNNNTSKSIPIGPRRAFYATGRIIGDIRQAYC	X4	35	S	R	5			X	
X4_3	6.0	225	CTRPNNNTKKGIAIGPGRTFYTREKIIGNIRQAHC	X4	35	G	K	6		X		X
R5_4	5.0	185	CTRPNNNTSKSIPIGPGRAWYATGRIIGDIRQAYC	R5	35	S	R	4				
X4_4	2.7	100	CTRPNNNTKKGIAIGPGRTFYTREKIIENIRQAHC	X4	35	G	K	5		X		
X4_5	1.2	44	CTRPNNNTRKGIAIGPGRTFYTREKIIGNIRLAHC	X4	35	G	K	6		X		X
R5_5	1.0	39	CTRPNNNTSKSIPIRPERAFYATGRIIGDIRQAYC	R5	35	S	R	4				
R5_6	1.0	38	CTRPNNNTRKSIPIGPGRAFYTRGEIIGNIRQAHC	R5	35	S	E	5				
R5_7	0.8	30	CTRPNNNTRKSITIGPGRAFYTRGEIIGNIRQAHC	R5	35	S	E	5				