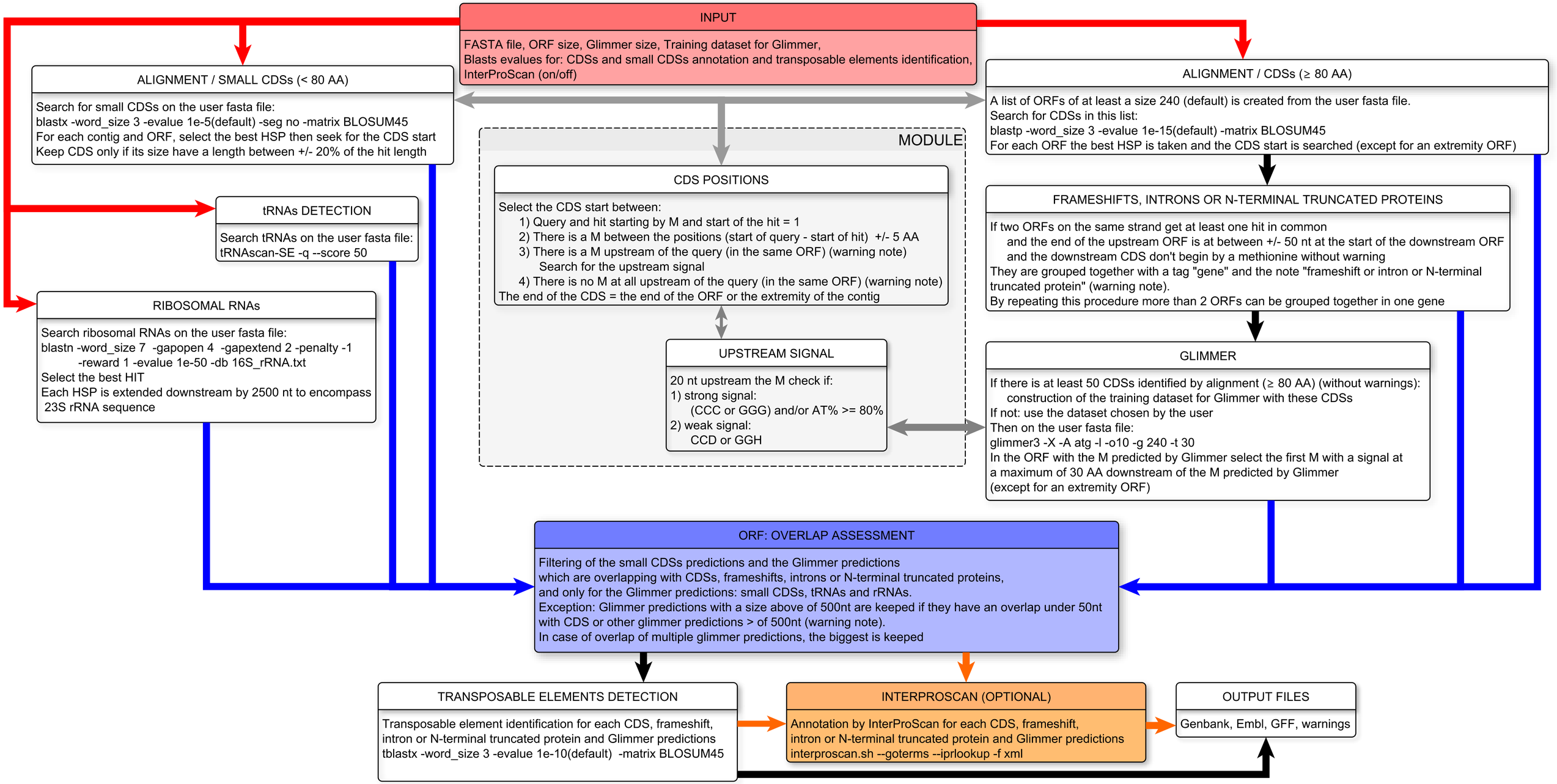
INPUT
FASTA file, ORF size, Glimmer size, Training dataset for Glimmer,
Blasts evalues for: CDSs and small CDSs annotation and transposable elements identification,
InterProScan (on/off)
ALIGNMENT / SMALL CDSs (< 80 AA)
Search for small CDSs on the user fasta file:
blastx -word_size 3 -evalue 1e-5(default) -seg no -matrix BLOSUM45
For each contig and ORF, select the best HSP then seek for the CDS start
Keep CDS only if its size have a length between +/- 20% of the hit length
ALIGNMENT / CDSs (≥ 80 AA)
A list of ORFs of at least a size 240 (default) is created from the user fasta file.
Search for CDSs in this list:
blastp -word_size 3 -evalue 1e-15(default) -matrix BLOSUM45
For each ORF the best HSP is taken and the CDS start is searched (except for an extremity ORF)
MODULE
CDS POSITIONS
Select the CDS start between:
1) Query and hit starting by M and start of the hit = 1
2) There is a M between the positions (start of query - start of hit) +/- 5 AA
3) There is a M upstream of the query (in the same ORF) (warning note)
Search for the upstream signal
4) There is no M at all upstream of the query (in the same ORF) (warning note)
The end of the CDS = the end of the ORF or the extremity of the contig
UPSTREAM SIGNAL
20 nt upstream the M check if:
1) strong signal:
(CCC or GGG) and/or AT% >= 80%
2) weak signal:
CCD or GGH
tRNAs DETECTION
Search tRNAs on the user fasta file:
tRNAscan-SE -q --score 50
RIBOSOMAL RNAs
Search ribosomal RNAs on the user fasta file:
blastn -word_size 7 -gapopen 4 -gapextend 2 -penalty -1
-reward 1 -evalue 1e-50 -db 16S_rRNA.txt
Select the best HIT
Each HSP is extended downstream by 2500 nt to encompass
23S rRNA sequence
FRAMESHIFTS, INTRONS OR N-TERMINAL TRUNCATED PROTEINS
If two ORFs on the same strand get at least one hit in common
and the end of the upstream ORF is at between +/- 50 nt at the start of the downstream ORF
and the downstream CDS don't begin by a methionine without warning
They are grouped together with a tag "gene" and the note "frameshift or intron or N-terminal truncated protein" (warning note).
By repeating this procedure more than 2 ORFs can be grouped together in one gene
GLIMMER
If there is at least 50 CDSs identified by alignment (≥ 80 AA) (without warnings):
construction of the training dataset for Glimmer with these CDSs
If not: use the dataset chosen by the user
Then on the user fasta file:
glimmer3 -X -A atg -l -o10 -g 240 -t 30
In the ORF with the M predicted by Glimmer select the first M with a signal at
a maximum of 30 AA downstream of the M predicted by Glimmer
(except for an extremity ORF)
ORF: OVERLAP ASSESSMENT
Filtering of the small CDSs predictions and the Glimmer predictions
which are overlapping with CDSs, frameshifts, introns or N-terminal truncated proteins,
and only for the Glimmer predictions: small CDSs, tRNAs and rRNAs.
Exception: Glimmer predictions with a size above of 500nt are keeped if they have an overlap under 50nt
with CDS or other glimmer predictions > of 500nt (warning note).
In case of overlap of multiple glimmer predictions, the biggest is keeped
TRANSPOSABLE ELEMENTS DETECTION
Transposable element identification for each CDS, frameshift,
intron or N-terminal truncated protein and Glimmer predictions
tblastx -word_size 3 -evalue 1e-10(default) -matrix BLOSUM45
INTERPROSCAN (OPTIONAL)
Annotation by InterProScan for each CDS, frameshift,
intron or N-terminal truncated protein and Glimmer predictions
interproscan.sh --goterms --iprlookup -f xml
OUTPUT FILES
Genbank, Embl, GFF, warnings