Ekenda
spider transcriptome analysis

Bioinformatic portal for spider venom gland transcriptome analysis

Improved Transcriptome Annotation Pipeline (iTAP)

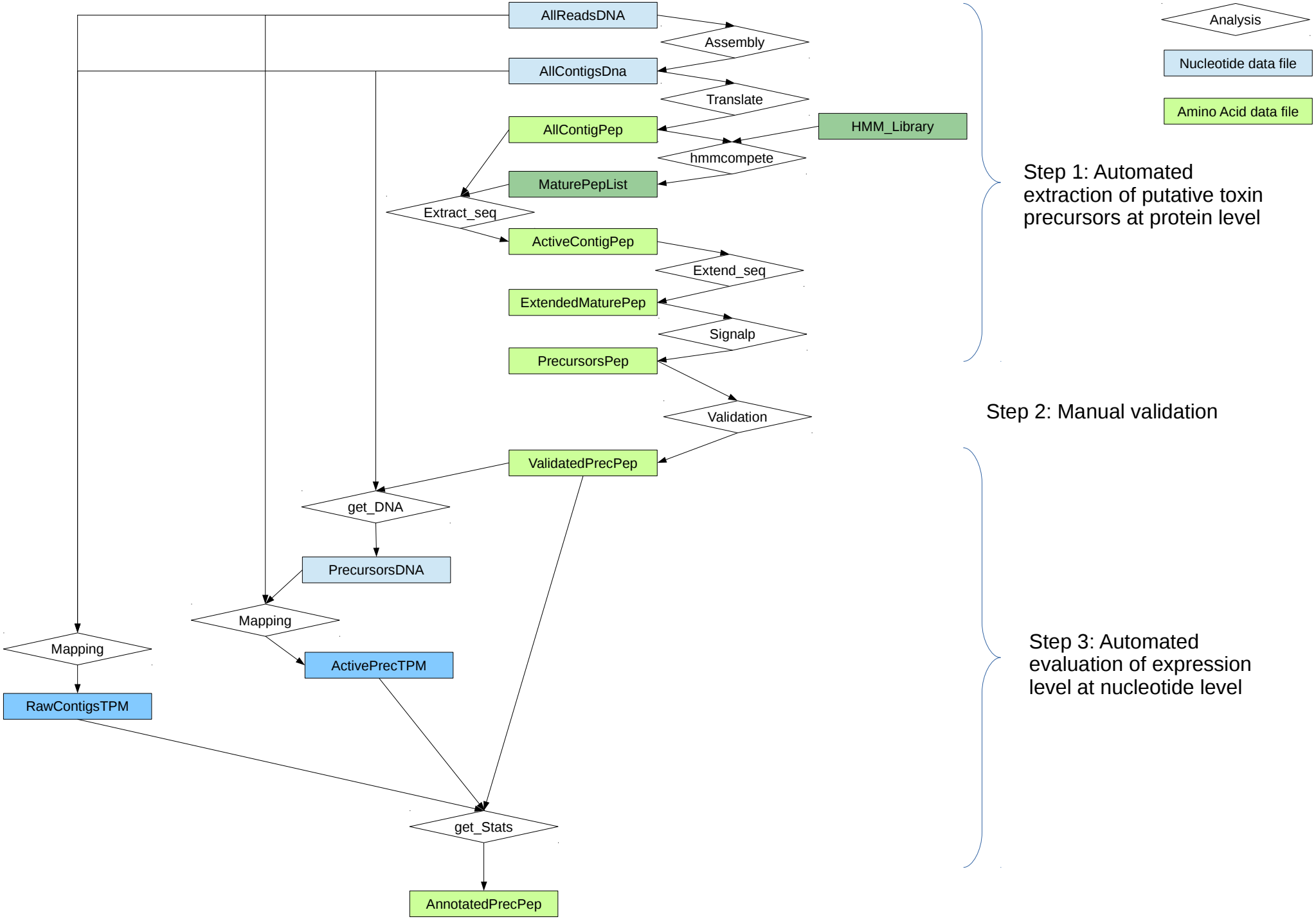
Dominique Koua
PhD in Bioinformatics

Aim of the project:

- Speed up transcriptome dataset analysis

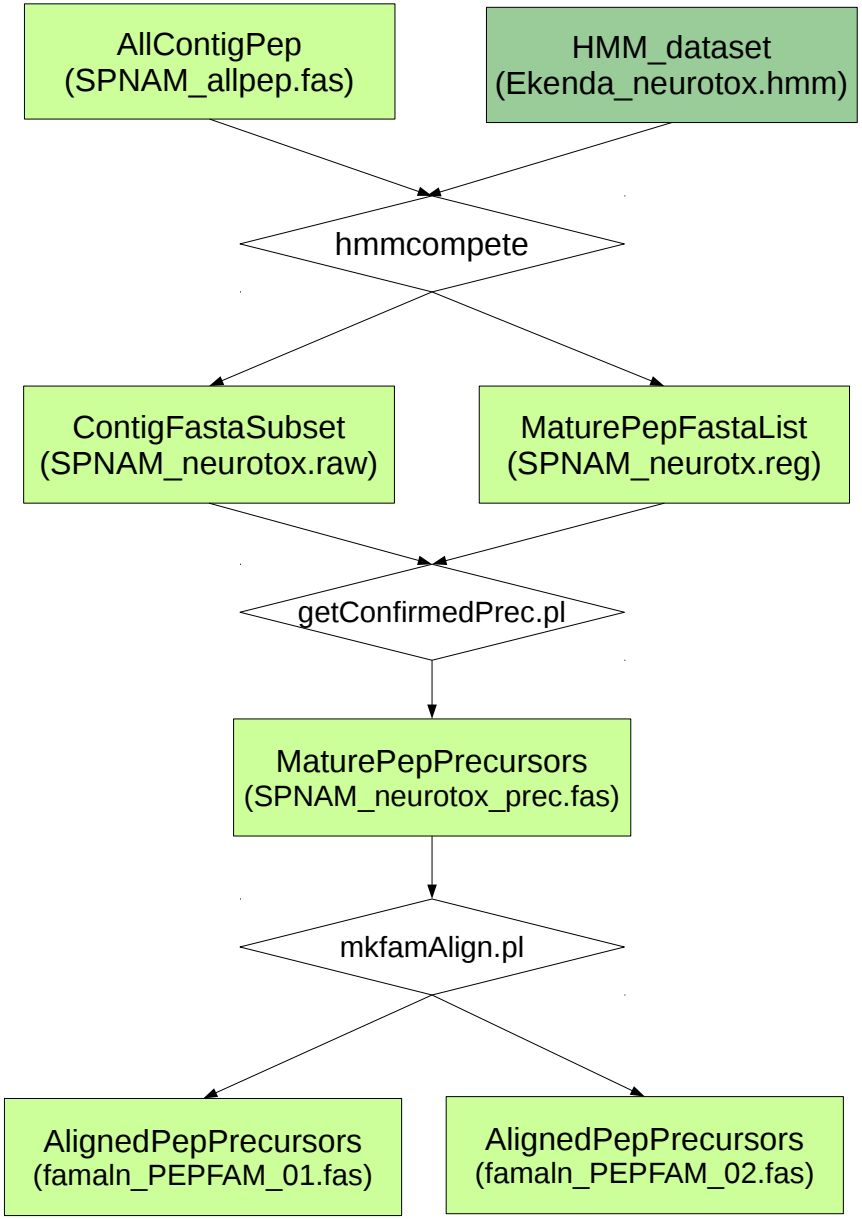
- Reliably extract peptides of interest

- Provide a easy-going method for users


Analysis
Nucleotide data file
Amino Acid data file
AllReadsDNA
Assembly
AllContigsDna
Translate
AllContigPep
HMM_Library
hmmcompete
MaturePepList
Extract_seq
ActiveContigPep
Extend_seq
ExtendedMaturePep
Signalp
PrecursorsPep
Validation
ValidatedPrecPep
get_DNA
PrecursorsDNA
Mapping
ActivePrecTPM
Mapping
RawContigsTPM
get_Stats
AnnotatedPrecPep
Step 1: Automated extraction of putative toxin precursors at protein level
Step 2: Manual validation
Step 3: Automated evaluation of expression level at nucleotide level

Detailed workflow and underlying programs (Step 1):

6-frames translation of assembled reads



Revised HMM models

New functionalities added to this specific purpose

Extract the most reliable between all possible precursors around the matched region (SignalP)

Separate possible precursors to facilitate human validation

Step 1 in real life:

1- Move to /home/data/assembly

2- For neurotoxin identification type:
sh pipeline_neurotox_step1.sh

For venom protein identification type:
sh pipeline_venprot_step1.sh

For others family:
- create the dedicated HMM
- copy the pipeline: cp pipeline_neurotox_step1.sh pipeline_myfam_step1.sh
- edit the new pipeline file: change all destination file names
- run the pipeline

Output of Step 1:

```
[root@IEESRVTOXIN LKN_00_CUPSA_454]# ls famaln_SN_*
famaln_SN_02_00.fas  famaln_SN_04_00.fas  famaln_SN_05_08.fas
famaln_SN_02_03.fas  famaln_SN_04_02.fas  famaln_SN_10_56.fas
famaln_SN_02_06.fas  famaln_SN_04_04.fas  famaln_SN_11_00.fas
famaln_SN_02_07.fas  famaln_SN_05_04.fas  famaln_SN_13_00.fas
famaln_SN_02_08.fas  famaln_SN_05_06.fas  famaln_SN_14_01.fas
famaln_SN_02_09.fas  famaln_SN_05_07.fas  famaln_SN_19_00.fas
```

Copy the files and open/edit with the preferred alignment program.

Or better:

Alignment of SN_02_03

```
>Contig_Spider_Gland_98_27070_1  #SN_02_03 Plectoxin superfamily, Type IV omega agatoxin family (IPR004169;)#   SIGNO
MLTPNQV-------------------------------KSRRNLDIAFRSGDDARGVTKCCAGRSCDCNVTRT------------
>Contig_Spider_Gland_98_20062_3  #SN_02_03 Plectoxin superfamily, Type IV omega agatoxin family (IPR004169;)#   SIGYES
---XSQVLLVLVGLIMFLGVHADTESSEITEESRYCIPKWRRC----------TWGGPKCCAGRSCDCNVTRTNCRCSPRLFGLG
>Contig_Spider_Gland_98_20129_2  #SN_02_03 Plectoxin superfamily, Type IV omega agatoxin family (IPR004169;)#   SIGYES
MWPKVQVLLVLVGLIMFLGVHADTESSEITEESRYCIPKWRRC----------TWGGPKCCAGRSCDCNVTRTNCRCSPRLLA--
>Contig_Spider_Gland_98_13888_5  #SN_02_03 Plectoxin superfamily, Type IV omega agatoxin family (IPR004169;)#   SIGYES
MWPKVQVLLVLVGLIMFLGVHADTESSEITEESRYCIPKWRRC----------TWGGPKCCAGRSCDCNVTRTNCRCSPRLFGLG
```

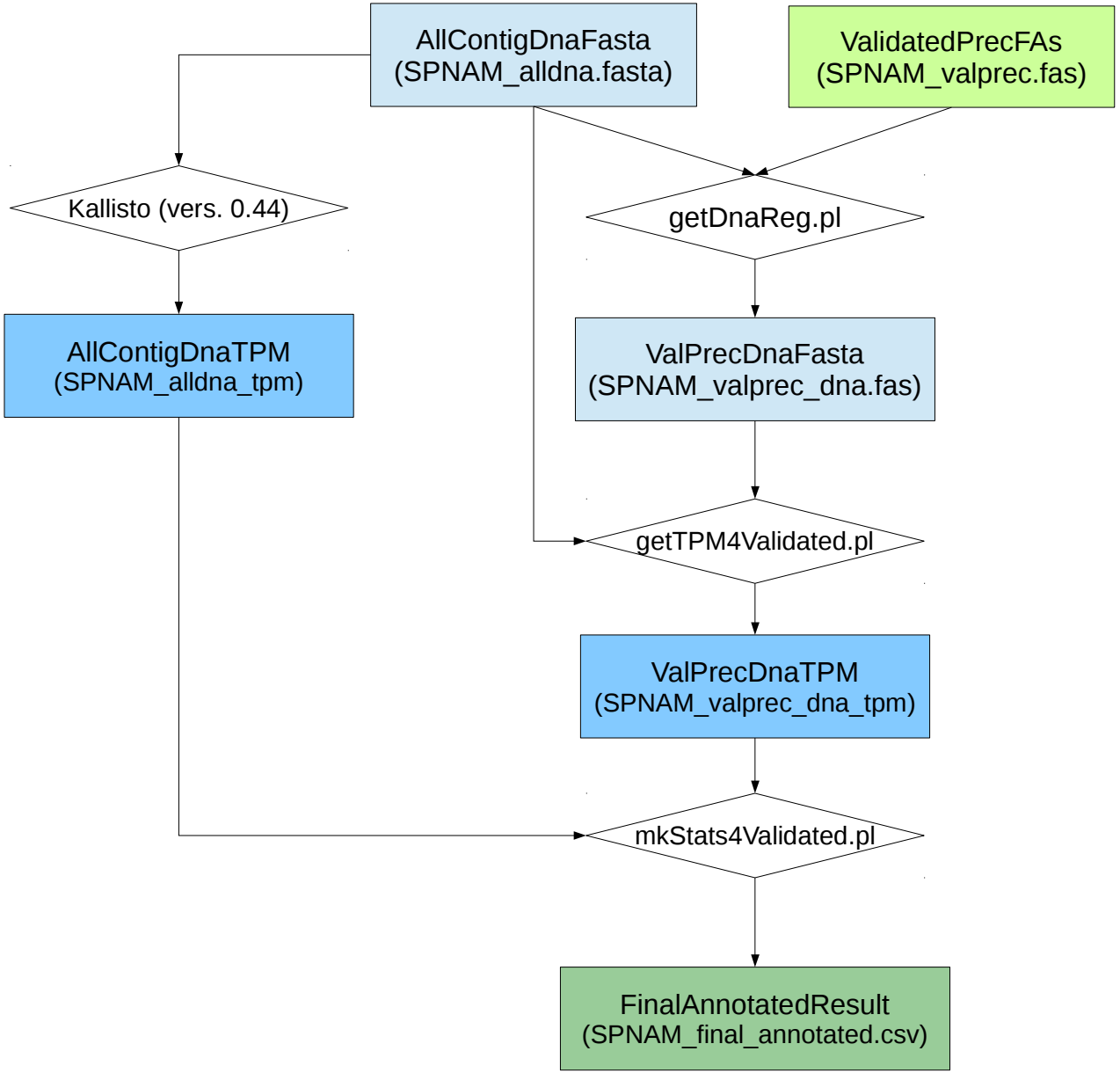
Alignment of SN_02_06

```
>Contig_Spider_Gland_98_31118_6  #SN_02_06 CsTx-36#   SIGYES
MRLLIPILMVVVAFIAVIGVHATAYSNENFENDPEGKRSCAEAYQTCDSIPCCNERSCVCNWLGKECKCKKSLGELIDTLLRILKIGKTSLNF
>Contig_Spider_Gland_98_13105_2  #SN_02_06 CsTx-36#   SIGYES
--------MVVVAFIAVIGVHATAYSNENFENDPEGNRSCAEAYQTCDSIPCCNERSCVCNWLRKR---------------------------
>Contig_Spider_Gland_98_14347_4  #SN_02_06 CsTx-36#   SIGYES
MRLLIPILMVVVAFIAVIGVHATAYSNDNFENDPEGNRSCAEAYQTCDSIPCCNERSCVCNWLGKECKCKKSLGELIDTLL-----GS-----
>Contig_Spider_Gland_98_16437_3  #SN_02_06 CsTx-36#   SIGYES
MRLLFPILMVVVAFIAVIGVHGTAYSNENFENDPEGNRSCAEAYQTCDSIPCCNERSCVCNWLGKDANARNHWEN------------------
>Contig_Spider_Gland_98_28964_4  #SN_02_06 CsTx-36#   SIGYES
--------MVVVAFIAVIGVHGTAYSNENFENDPEGNRSCAEAYQTCDSIPCCNERSCVCNWLGKRMQMQ----------------------
>Contig_Spider_Gland_98_14984_2  #SN_02_06 CsTx-36#   SIGYES
--------MVVVAFIAVIGVHGTAYSNENFENDPEGNRSCAEAYQTCDSIPCCNERSCVCNWLGKECKCKKSLGELIDTLL-----GS-----
>Contig_Spider_Gland_98_9893_5  #SN_02_06 CsTx-36#   SIGYES
--------MVVVAFIAVIGVHATAYSNDNFENDPEGNRSCAEAYQTCDSIPCCNERSCVCNWLGKECKCKKSLGELIDTLL-----GS-----
```

Directly load annotations on the Ekenda platform. Copy and edit there.

Note the family as well as the ‘SIGYES/SIGNO’ annotations

Detailed workflow and underlying programs (Step 2):



Manually merge all validated alignment files

Extract DNA region

Independent run of kalisto

Report Sequence_id, Family, TPMs and aaSequence

Step 2 in real life:

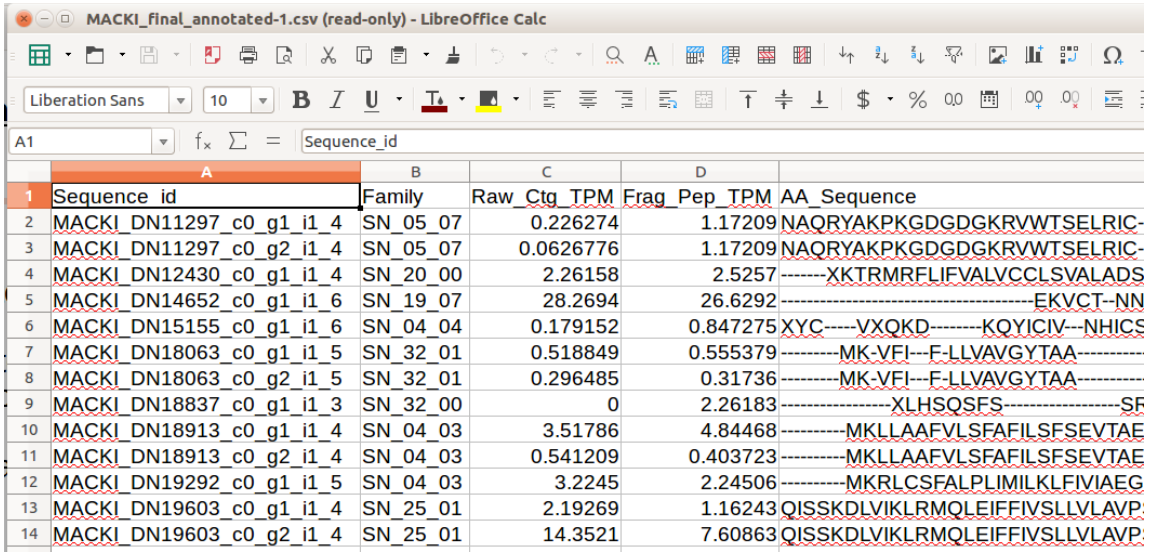
1- Move to /home/data/assembly

2- Create a single file for all validated peptides: name must be like SPNAM_valprec.fas

3- run step 2:
    sh pipeline_all_step2.sh

Your final result will be in a file named : SPNAM_final_annotated.csv

The validated peptide file will be renamed to SPNAM_valprec_done.fas



MACKI_final_annotated-1.csv (read-only) - LibreOffice Calc

	A	B	C	D	
1	Sequence_id	Family	Raw_Ctg_TPM	Frag_Pep_TPM	AA_Sequence
2	MACKI_DN11297_c0_g1_i1_4	SN_05_07	0.226274	1.17209	NAQRYAKPKGDGDGKRVWTSELRIC-
3	MACKI_DN11297_c0_g2_i1_4	SN_05_07	0.0626776	1.17209	NAQRYAKPKGDGDGKRVWTSELRIC-
4	MACKI_DN12430_c0_g1_i1_4	SN_20_00	2.26158	2.5257	-------XKTRMRFLIFVALVCCLSVALADS
5	MACKI_DN14652_c0_g1_i1_6	SN_19_07	28.2694	26.6292	----------EKVCT--NN
6	MACKI_DN15155_c0_g1_i1_6	SN_04_04	0.179152	0.847275	XYC-----VXQKD--------KQYICIV---NHICS
7	MACKI_DN18063_c0_g1_i1_5	SN_32_01	0.518849	0.555379	---------MK-VFI---F-LLVAVGYTAA----------
8	MACKI_DN18063_c0_g2_i1_5	SN_32_01	0.296485	0.31736	---------MK-VFI---F-LLVAVGYTAA----------
9	MACKI_DN18837_c0_g1_i1_3	SN_32_00	0	2.26183	----------XLHSQSFS----------SF
10	MACKI_DN18913_c0_g1_i1_4	SN_04_03	3.51786	4.84468	----------MKLLAAFVLSFAFILSFSEVTAE
11	MACKI_DN18913_c0_g2_i1_4	SN_04_03	0.541209	0.403723	----------MKLLAAFVLSFAFILSFSEVTAE
12	MACKI_DN19292_c0_g1_i1_5	SN_04_03	3.2245	2.24506	----------MKRLCSFALPLIMILKLFIVIAEG
13	MACKI_DN19603_c0_g1_i1_4	SN_25_01	2.19269	1.16243	QISSKDLVIKLRMQLEIFFIVSLLVLAVP
14	MACKI_DN19603_c0_g2_i1_4	SN_25_01	14.3521	7.60863	QISSKDLVIKLRMQLEIFFIVSLLVLAVP