Global Analysis of *Leishmania* genes expression using SAGE Libraries

Sondos SMANDI
LIVGM

Bioinformatics for Africa,
Nairobi 2007

OUTLINE

1) *Leishmania*

2) Objectives

3) TAGs assignment

4) Differential expression

5) Conclusion

6) Perspectives

1) *Leishmania*

- *Leishmania* is flagellated protozoan parasite belonging to the order of *Kinetoplastida* and to the family of *Trypanosomatidae*
- Transmitted by sandflies
- It alternates between an amastigote stage in macrophage and a promastigote stage in the digestive tract of sandflies
- It is responsible of leishmaniasis, a parasitic disease for which there is no vaccine and existing drugs are toxic

1) *Leishmania*

Genome structure and *content*

- The *leishmania major* genome is about 33 Mb organized in 36 Chromosomes
- 8272 protein coding genes
- Organized into 133 clusters of genes
- GC content is 59.7%

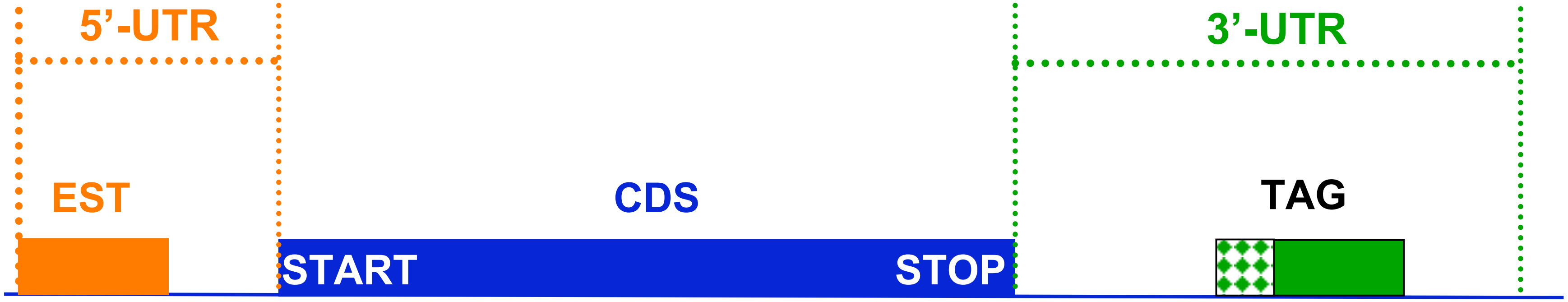
Transcription

- Transcription is polycistronic (genes are grouped in large units of transcription)
- The protein coding genes are almost never interrupted by introns
- The mature mRNAs are generated from primary transcripts by trans-splicing and polyadenylation

2) Objectives

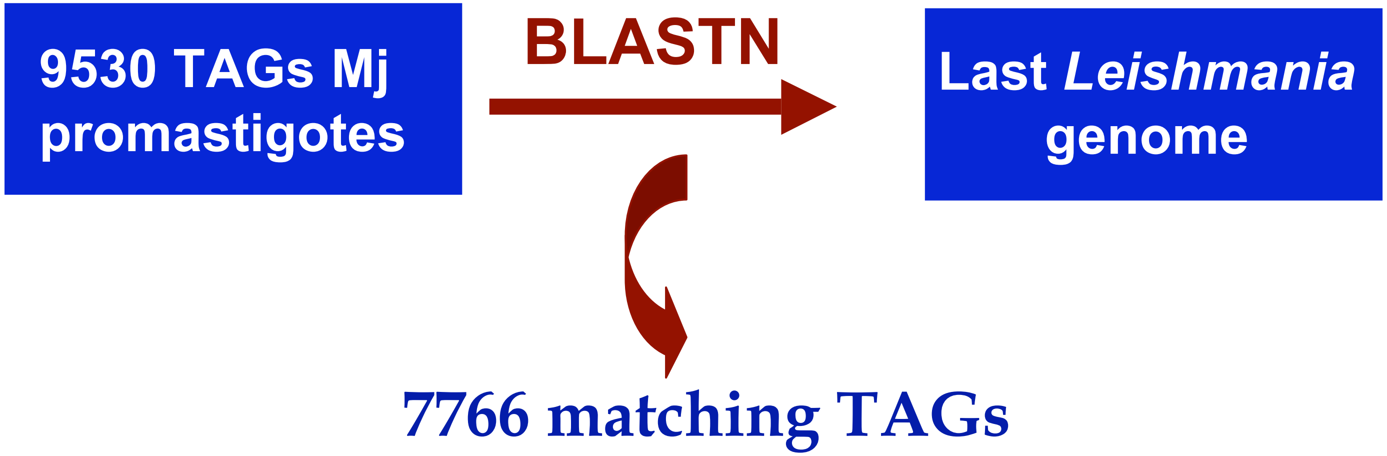
- In order to study the impact of intracellular infection on the expression of macrophage genes, LIVGM has developed three SAGE libraries:
    - Non-Infected Macrophages (NS, 32332 TAGs, 13938 unique TAGs)
    - *Leishmania* promastigote (Mj, 33906 TAGs, 9530 unique TAGs)
    - Macrophages Infected by *Leishmania major* (amastigote) (Lm, 62136 TAGs, 24418 unique TAGs)
- To analyse the latter two libraries, and translate them into more comprehensive and functional information, it was a necessary to:
    - Locate the *Leishmania* TAGs on *Leishmania* genome
    - Assign TAG to corresponding gene
    - Evaluate the parasite genes expression

2) Objectives



- Very few full length *Leishmania* cDNA are available
- The only *Leishmania* genes evidences are the CDS predictions available in GeneDB
- The TAGs are normally located in 3'UTR (outside the CDS)

3) TAGs assignment



9530 TAGs Mj promastigotes

BLASTN

Last *Leishmania* genome

7766 matching TAGs

The First step was to map all TAGs to the parasite genome by BLAST against the last parasite genome release (downloaded from www.genedb.org)

3) TAGs assignment

- Then we mapped other data that could have been useful for addressing our problem:
  - All *L. major* ESTs,
  - TAGs from other *Leishmania,*
  - All CATG.
- All these mappings were very helpful to identify the strategy to follow.

3) TAGs assignment

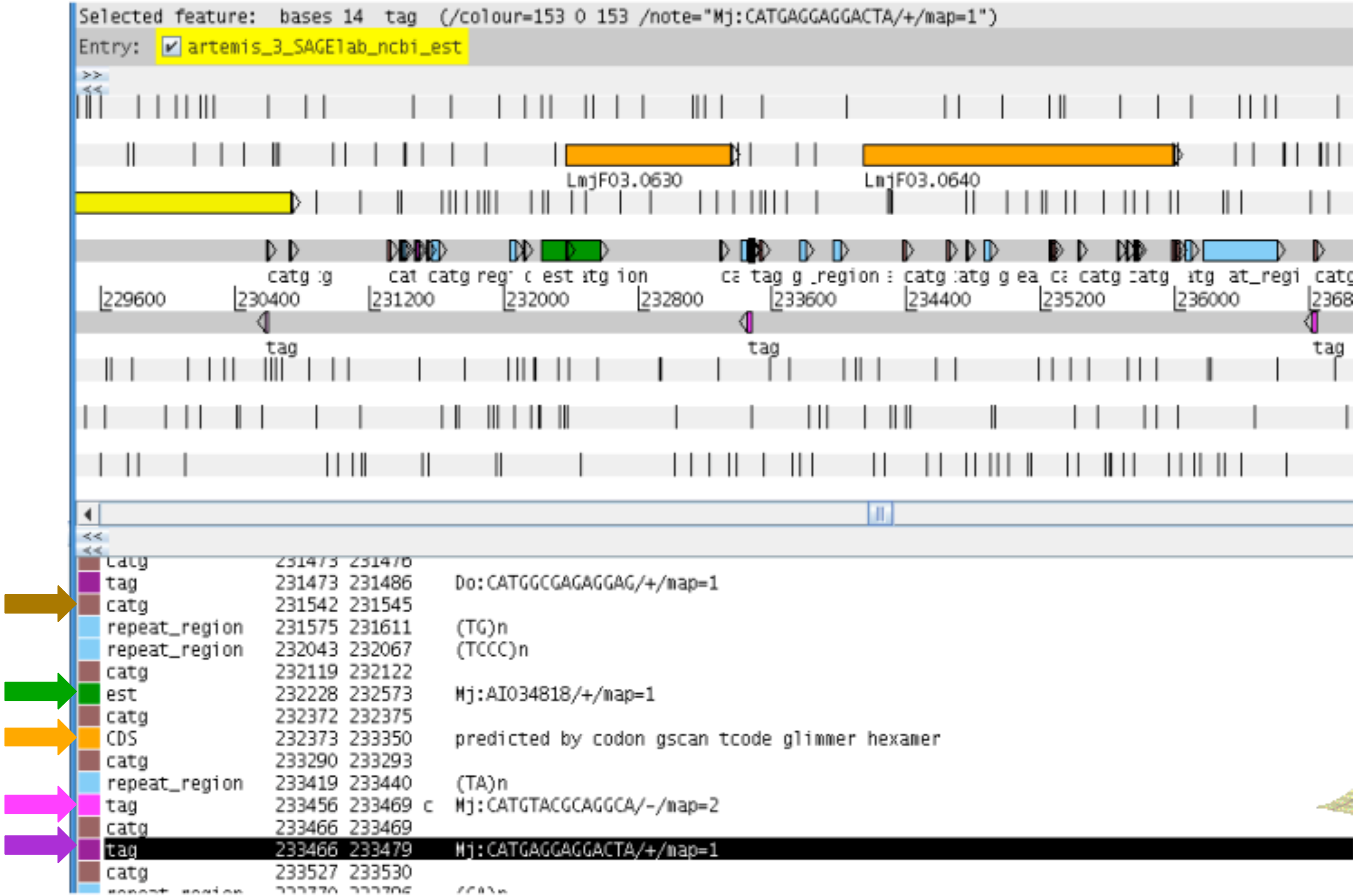
Integration and visualisation of data

- Configure and Integrate all mapped data into the ARTEMIS files (downloaded from GeneDB)
- Each data type (single match, double match...) was labelled with a colour
- Using ARTEMIS for mapped data visualization (Rutherford et al, 2000)

Results of mapping TAGs and ESTs

Count	Dataset	Color code
7766 Matched TAGs		
4191	One match	🟪
1919	Two matches	🟥
1656	More than two matches	⬜
1567 Matched ESTs		
1167	One match	🟩
336	Two matches	🟩
73	More than two matches	⬜

143862	catg	🟫



Selected feature: bases 14 tag (/colour=153 0 153 /note="Mj:CATGAGGAGGACTA/+/map=1")

Entry: ☑ artemis_3_SAGElab_ncbi_est

```
catg            231473 231476
tag             231473 231486     Do:CATGGCGAGAGGAG/+/map=1
catg            231542 231545
repeat_region   231575 231611     (TG)n
repeat_region   232043 232067     (TCCC)n
catg            232119 232122
est             232228 232573     Mj:AI034818/+/map=1
catg            232372 232375
CDS             232373 233350     predicted by codon gscan tcode glimmer hexamer
catg            233290 233293
repeat_region   233419 233440     (TA)n
tag             233456 233469 c   Mj:CATGTACGCAGGCA/-/map=2
catg            233466 233469
tag             233466 233479     Mj:CATGAGGAGGACTA/+/map=1
catg            233527 233530
repeat_region   233770 233796     (CA)n
```

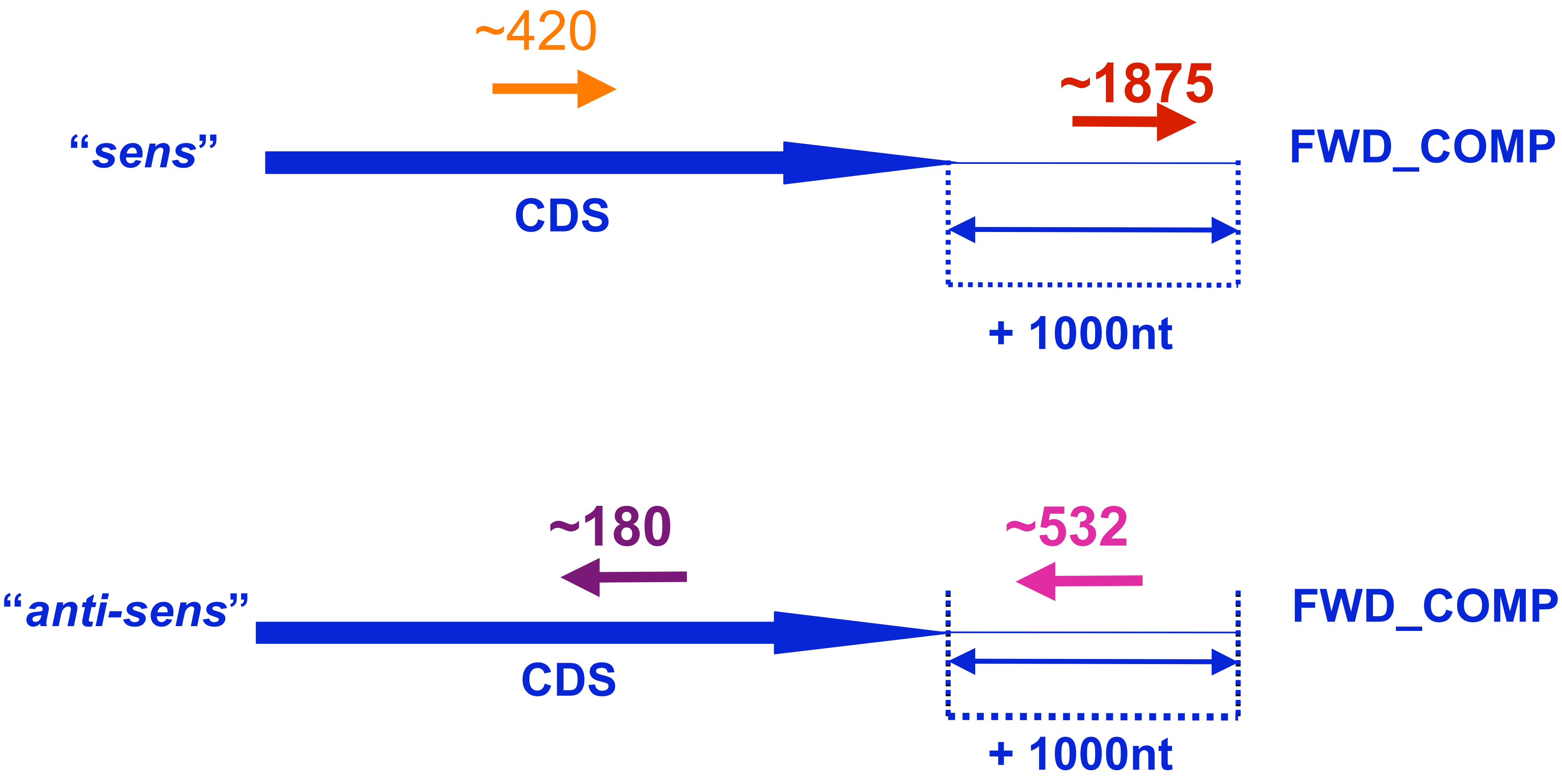
ARTEMIS was also useful in debugging the developed programs

3) TAGs assignment

Assignment of single match TAGs

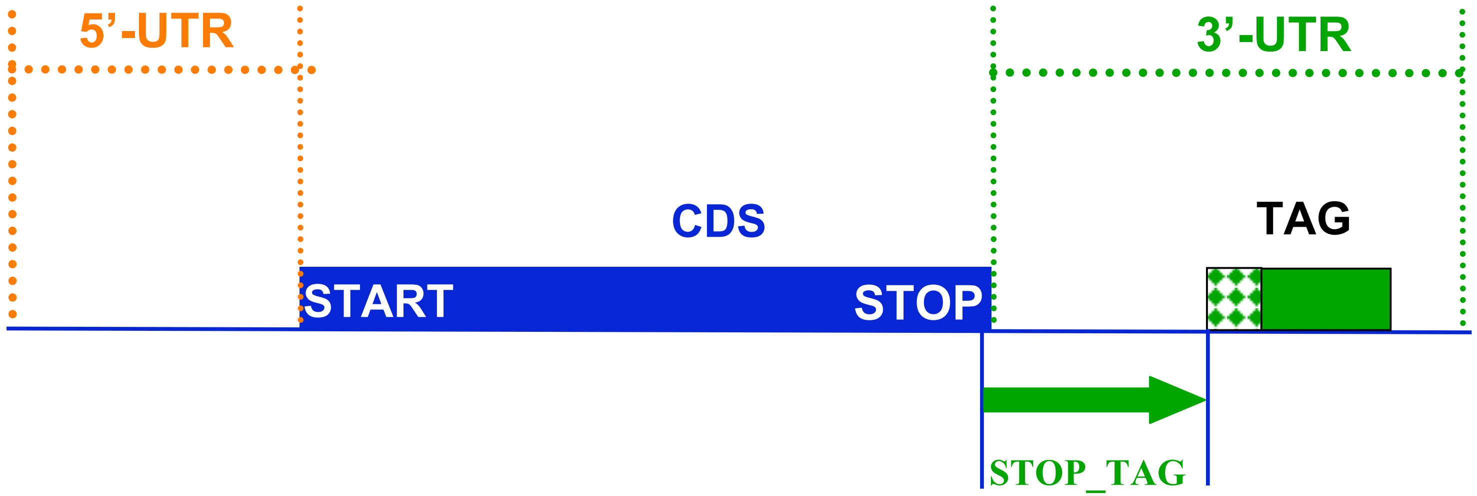
- Using the *Leishmania* gene catalogue and what is known about the size of the 3’UTR of *Leishmania* genes, we were able to easily assign 3007 TAGs:
    - TAGs that are inside the genes and in the same direction (“*sens*”)
    - TAGs that are inside the genes and in the opposite direction (“*anti-sens*”)
    - TAGs that are in the 1000nt that follow the STOP of genes (3'UTR) and in the same direction (“*sens*”)
    - TAGs that are in the 1000nt that follow the STOP of genes (3'UTR) and in the opposite direction (“*anti-sens*”)

Results: Assignment of single match TAGs



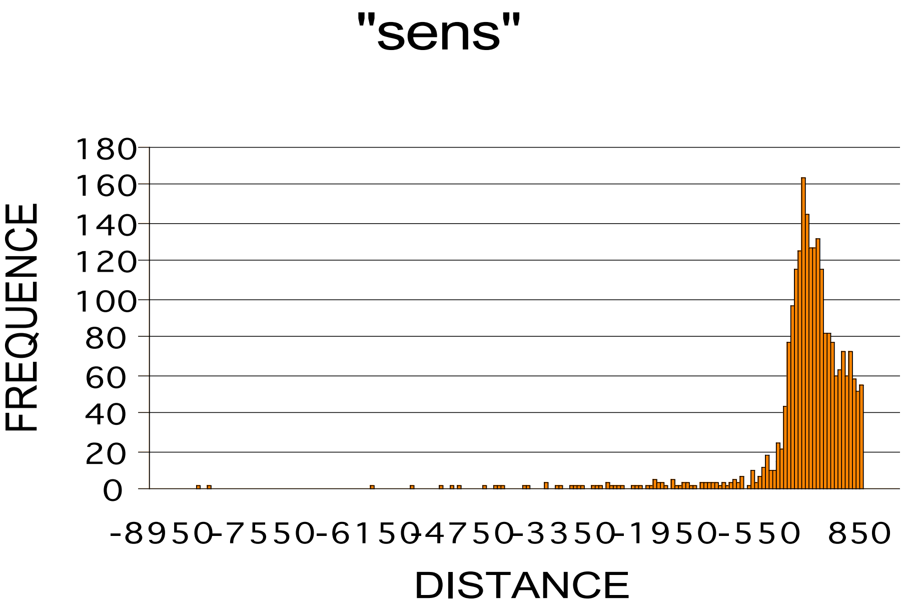
3) TAGs assignment

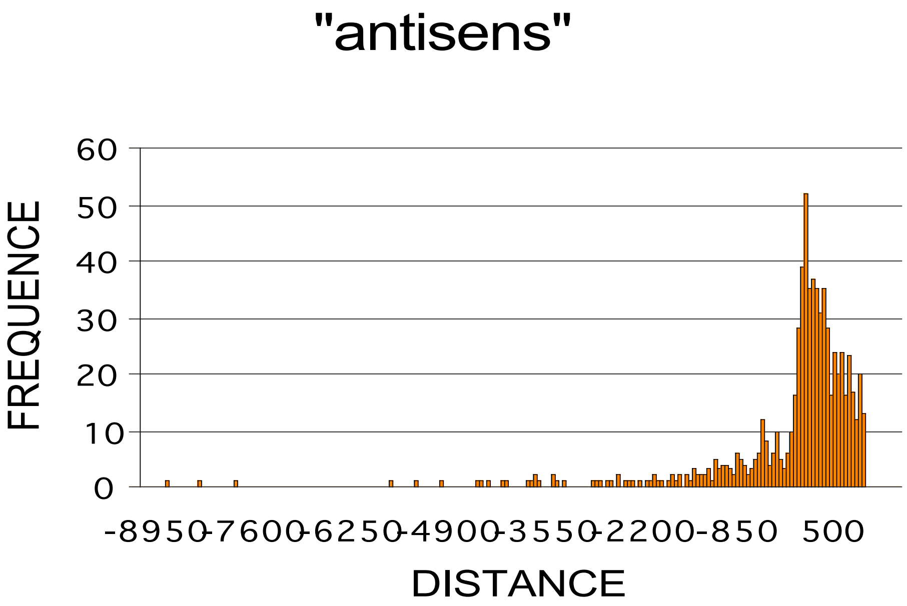
Estimation of distance "STOP-TAG" for the assigned 3007 TAGs



Extract the distance between each TAG and the STOP codon of the corresponding gene

3) TAGs assignment





- These histograms illustrate the distribution of the recorded distances between the TAGs and the closest STOP (gene).

  Population="assigned TAGs", **Size**=3007

- We need to extract out of this sample a statistic that reflects the behaviour/ distribution of the recorded distance

3) TAGs assignment

- We used the Gaussian Kernels in order to estimate the «density distribution» of the distances «STOP_TAG».
- The formula of the Gaussian Kernels follows:

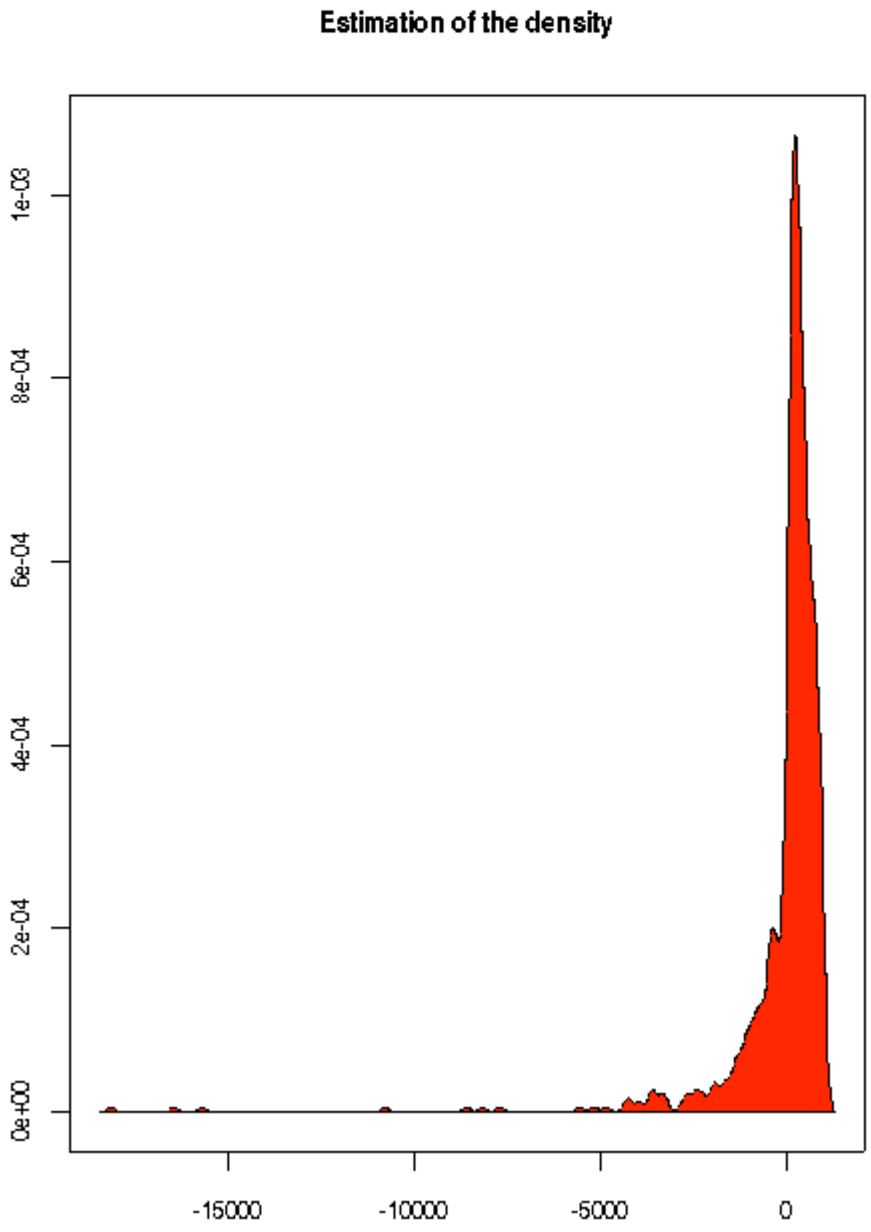
$$f(x) = \frac{1}{nh} \sum_{i,\, x_i \in V(x)} K\left(\frac{x_i - x}{h}\right)$$

H: is the bandwidth (estimated from data in order to minimize SSE,

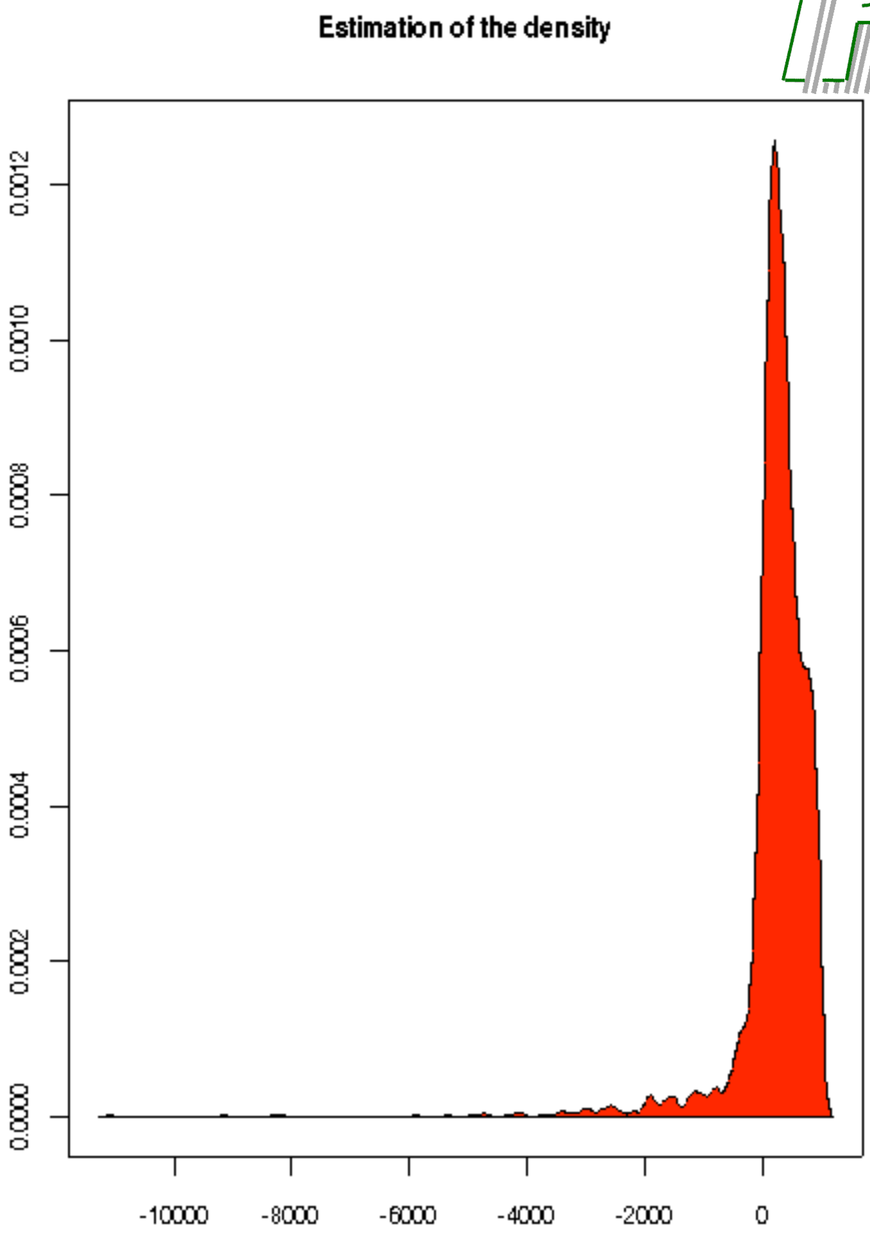
N: sample size

K: a standard Gaussian normal distribution.

- The estimated density will be used to assign the multiple match TAGs to their



« anti-sens »



« sens »

3) TAGs assignment

Assignment of the multiple match TAGs

- 13617 “mappings” correspond to the 3575 TAGs with multiple matches
- The distance between these “mappings” and the STOP of the nearest gene has been extracted
- The most likely distances have been kept

$P(x \leq \text{dist} < x+50)$ close to 1, means that the distance between the TAG and the STOP is likely to correspond to a distance between the TAG and the gene it belongs to

➡ This TAG can probably be assigned to this gene

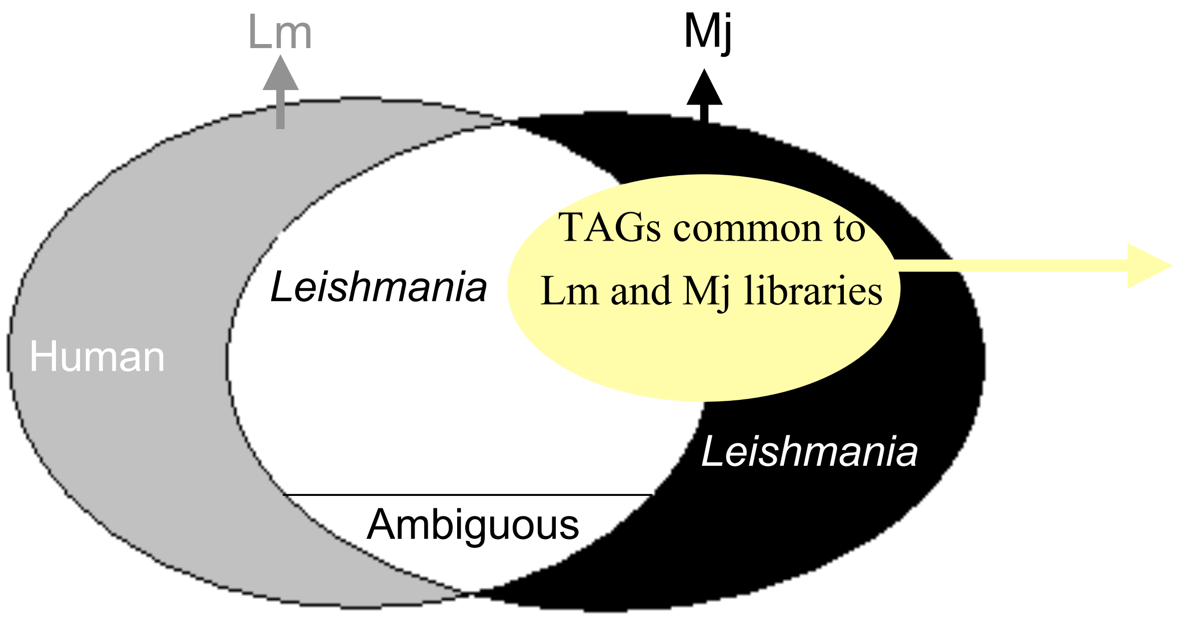
These statistics allowed to assign 1315 out of the 3575 TAGs with multiple matches ($P>0.01$)

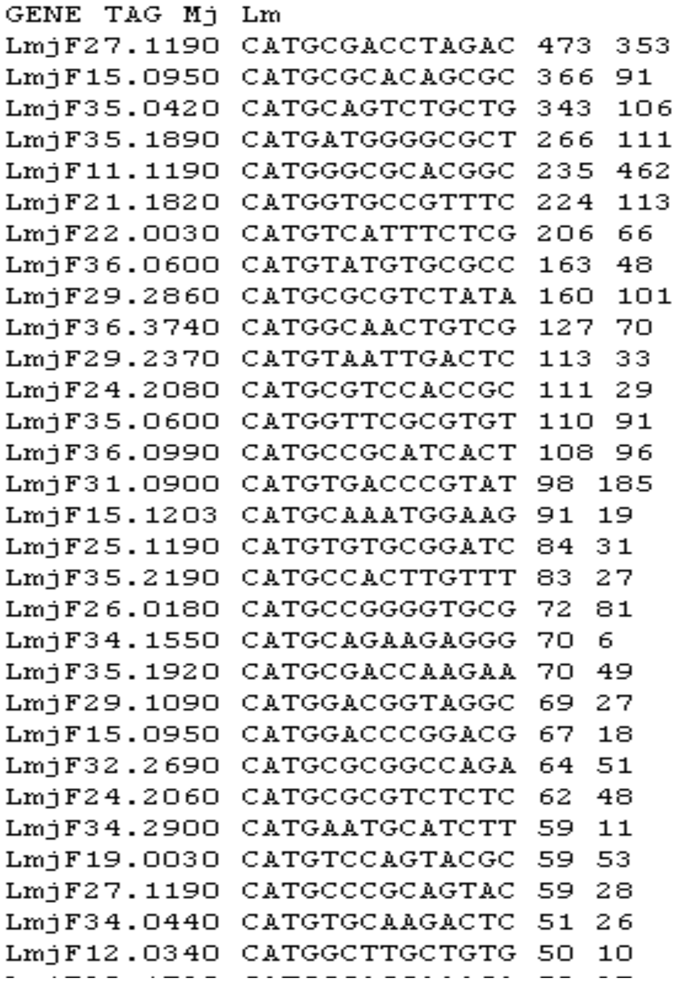
3) TAGs assignment: Results

- 9530 TAGs in SAGE library (*Leishmania* promastigote)
- 7766 matching TAGs (a single or a multiple match)
- 4322 TAGs were assigned to the corresponding genes

4) Differential expression

Evaluation of the expression of *Leishmania* genes





GENE	TAG	Mj	Lm
LmjF27.1190	CATGCGACCTAGAC	473	353
LmjF15.0950	CATGCGCACAGCGC	366	91
LmjF35.0420	CATGCAGTCTGCTG	343	106
LmjF35.1890	CATGATGGGGCGCT	266	111
LmjF11.1190	CATGGGCGCACGGC	235	462
LmjF21.1820	CATGGTGCCGTTTC	224	113
LmjF22.0030	CATGTCATTTCTCG	206	66
LmjF36.0600	CATGTATGTGCGCC	163	48
LmjF29.2860	CATGCGCGTCTATA	160	101
LmjF36.3740	CATGGCAACTGTCG	127	70
LmjF29.2370	CATGTAATTGACTC	113	33
LmjF24.2080	CATGCGTCCACCGC	111	29
LmjF35.0600	CATGGTTCGCGTGT	110	91
LmjF36.0990	CATGCCGCATCACT	108	96
LmjF31.0900	CATGTGACCCGTAT	98	185
LmjF15.1203	CATGCAAATGGAAG	91	19
LmjF25.1190	CATGTGTGCGGATC	84	31
LmjF35.2190	CATGCCACTTGTTT	83	27
LmjF26.0180	CATGCCGGGGTGCG	72	81
LmjF34.1550	CATGCAGAAGAGGG	70	6
LmjF35.1920	CATGCGACCAAGAA	70	49
LmjF29.1090	CATGGACGGTAGGC	69	27
LmjF15.0950	CATGGACCCGGACG	67	18
LmjF32.2690	CATGCGCGGCCAGA	64	51
LmjF24.2060	CATGCGCGTCTCTC	62	48
LmjF34.2900	CATGAATGCATCTT	59	11
LmjF19.0030	CATGTCCAGTACGC	59	53
LmjF27.1190	CATGCCCGCAGTAC	59	28
LmjF34.0440	CATGTGCAAGACTC	51	26
LmjF12.0340	CATGGCTTGCTGTG	50	10

- 1778 TAGs have been kept to evaluate the level of expression of the corresponding genes

4) Differential expression

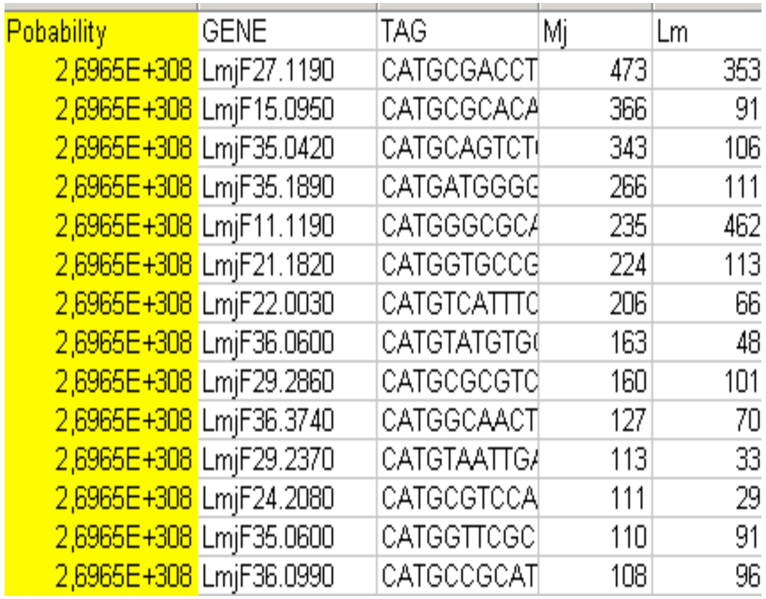
Evaluation of the expression of *Leishmania* genes

- We applied a probability based on the binomial distribution, used by our collaborators from Skuld-Tech (France), for the identification of differentially expressed TAGs

- P close to "**0**" means that there is a biologically meaningful expression variation → genes are differentially expressed
- P close to "**1**" means that the expression difference is not biologically significant → genes are not differentially expressed

4) Differential expression: Results

94 genes differentially expressed (P ≤ 0.01)

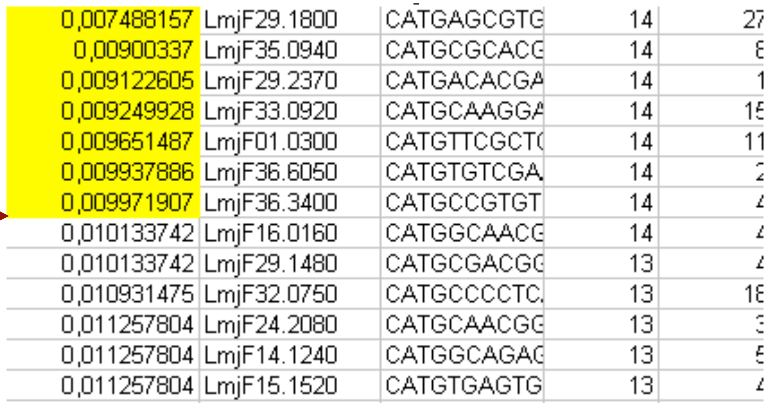


Pobability	GENE	TAG	Mj	Lm
2,6965E+308	LmjF27.1190	CATGCGACCT	473	353
2,6965E+308	LmjF15.0950	CATGCGCACA	366	91
2,6965E+308	LmjF35.0420	CATGCAGTCT	343	106
2,6965E+308	LmjF35.1890	CATGATGGGG	266	111
2,6965E+308	LmjF11.1190	CATGGGCGCA	235	462
2,6965E+308	LmjF21.1820	CATGGTGCCG	224	113
2,6965E+308	LmjF22.0030	CATGTCATTTC	206	66
2,6965E+308	LmjF36.0600	CATGTATGTG	163	48
2,6965E+308	LmjF29.2860	CATGCGCGTC	160	101
2,6965E+308	LmjF36.3740	CATGGCAACT	127	70
2,6965E+308	LmjF29.2370	CATGTAATTGA	113	33
2,6965E+308	LmjF24.2080	CATGCGTCCA	111	29
2,6965E+308	LmjF35.0600	CATGGTTCGC	110	91
2,6965E+308	LmjF36.0990	CATGCCGCAT	108	96

≤

0.01 →

<



0,007488157	LmjF29.1800	CATGAGCGTG	14	27
0,00900337	LmjF35.0940	CATGCGCACG	14	8
0,009122605	LmjF29.2370	CATGACACGA	14	1
0,009249928	LmjF33.0920	CATGCAAGGA	14	15
0,009651487	LmjF01.0300	CATGTTCGCT	14	11
0,009937886	LmjF36.6050	CATGTGTCGA	14	2
0,009971907	LmjF36.3400	CATGCCGTGT	14	4
0,010133742	LmjF16.0160	CATGGCAACG	14	4
0,010133742	LmjF29.1480	CATGCGACGG	13	4
0,010931475	LmjF32.0750	CATGCCCCTC	13	18
0,011257804	LmjF24.2080	CATGCAACGG	13	3
0,011257804	LmjF14.1240	CATGGCAGAG	13	5
0,011257804	LmjF15.1520	CATGTGAGTG	13	4

4) Conclusion

1. Assign *Leishmania* TAGs to the corresponding gene

Gaussian Kernels

Successful for 4305/9530 TAGs

1184 single match TAGs, 2260 multiple match TAGs and 1764 no match TAGs (5208 TAGs)are still not assigned:

- Some are due to sequencing errors (false positives) or due to the genetic variations between the strains.
- Some belong to new genes absent from the catalogue.
- The multiple match TAGs exist everywhere in the genome

2. Identify differentially expressed genes

Binomial distribution

94 genes are differentially expressed

5) Perspectives

- Increasing the number of assigned TAGs
- Clustering the *Leishmania* genes using their expression level
- Analyzing functionally these genes

THANK YOU