UTR motif discovery to investigate differential protein expression in *Leishmanias*

Carol L Farris

Level of relevant detail within a basic science publication has changed over time

JOURNAL OF BACTERIOLOGY, Dec. 1992, p. 8016–8022
0021-9193/92/248016-07$02.00/0

Vol. 174, No. 24

Dec 1992

Identification, Cloning, and Characterization of *rcsF*, a New Regulator Gene for Exopolysaccharide Synthesis That Suppresses the Division Mutation *ftsZ84* in *Escherichia coli* K-12

FRANÇOIS G. GERVAIS AND GABRIEL R. DRAPEAU*
Department of Microbiology, Université de Montréal, Montréal, Québec, Canada H3C 3J7

Received 8 June 1992/Accepted 15 October 1992

Result: Gene found, when deleted makes bacteria slimy

JOURNAL OF BACTERIOLOGY, Oct. 2003, p. 5735–5746
0021-9193/03/$08.00+0 DOI: 10.1128/JB.185.19.5735–5746.2003

Vol. 185, No. 19

Oct 2003

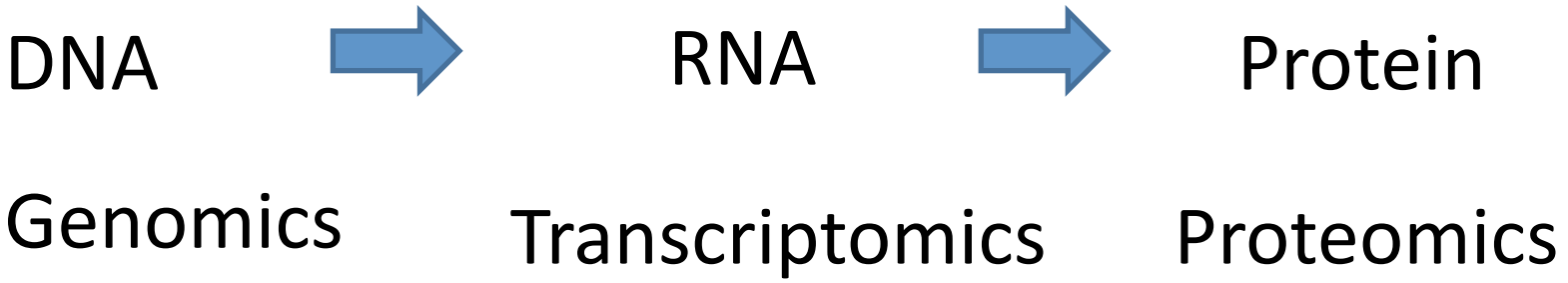
Genome-Wide Analyses Revealing a Signaling Network of the RcsC-YojN-RcsB Phosphorelay System in *Escherichia coli*

Daisuke Hagiwara,[1] Masahito Sugiura,[1] Taku Oshima,[2] Hirotada Mori,[2] Hirofumi Aiba,[1] Takafumi Yamashino,[1] and Takeshi Mizuno[1]*
Laboratory of Molecular Microbiology, School of Agriculture, Nagoya University, Chikusa-ku, Nagoya 464-8601,[1] and Research and Education Center for Genetic Information, Nara Institute of Science and Technology, Ikoma 630-0101,[2] Japan

Received 12 May 2003/Accepted 10 July 2003

Result: Gene incites signaling cascade that upregulates over 50 genes when exposed to nasty chemicals. Many don't appear to be involved in making bacteria slimy.

High throughput technologies have advanced basic science to systems level studies of an organism



Opportunities and Challenges for data management in the basic sciences

- Opportunities:
  - Large scale analysis of genomic, proteomic and transcriptomic data sets
  - More complete picture of organism behavior
    - details for a single gene & details on a genome wide scale
  - Redundant information reporting (supports previous conclusions)

- Challenges
  - Redundant information reporting (duplication of effort)
  - How to sift through all of the information?
    - Are we missing something?
    - Do we have an accurate picture?
  - Large datasets only have selective analysis on low hanging fruit
    - Top most significant hits discussed
    - Only gene or gene groups that are the focus of the lab are discussed
    - Results that support the paradigm of the research group are emphasized.

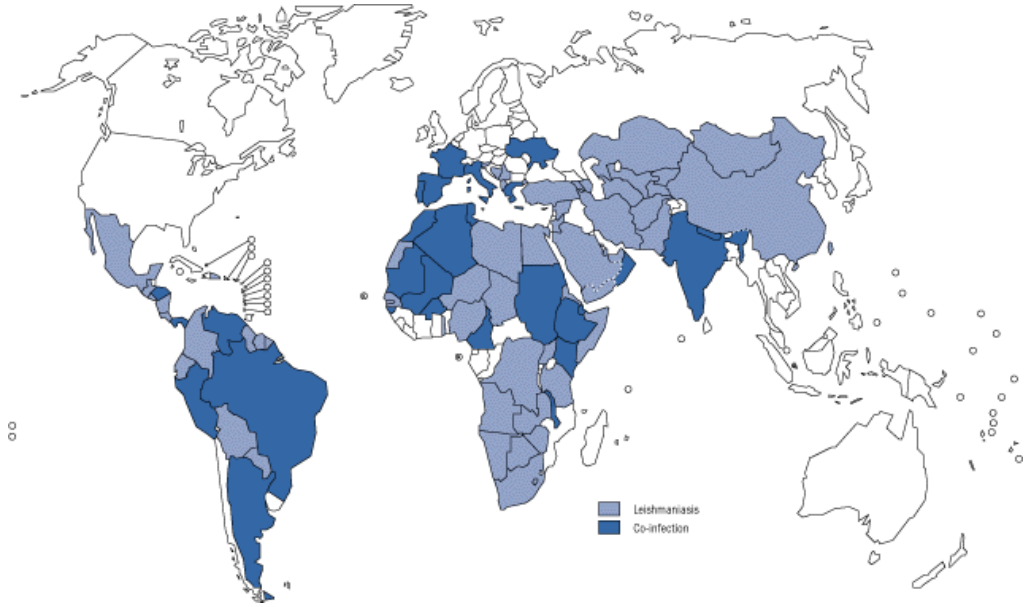
Finding the needles in the haystack: Assembling novel information from published data to direct research efforts at the bench

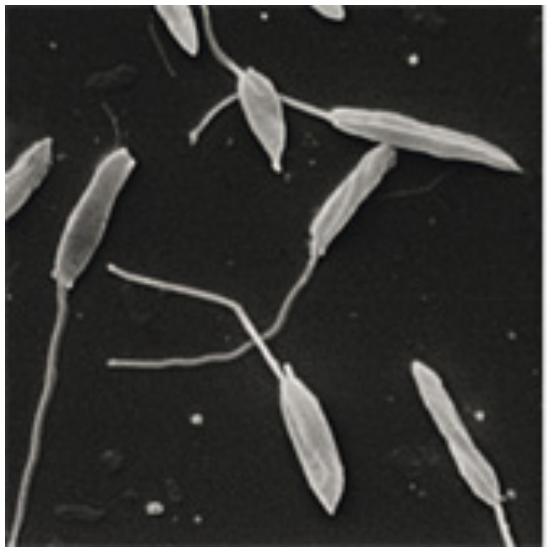
Project focus: Make use of advances in text mining, pattern matching to answer key questions about “difficult” pathogens.

…and *Leishmania* are "difficult" pathogens

PLoS NEGLECTED TROPICAL DISEASES

a peer-reviewed open-access journal published by the Public Library of Science

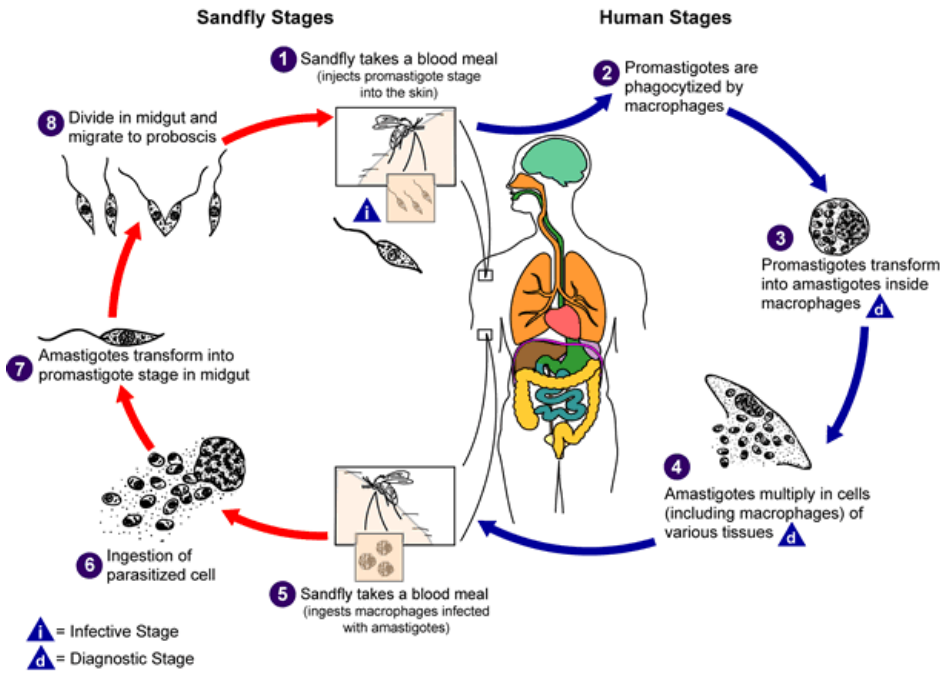




L. mexicana

Leishmania pathogenesis

Sandfly



Leishmania have two life stages:

1. Insect life stage (promastigotes)
2. Human life stage (amastigotes)

Human Host

Leishmaniasis:

<u>Cutaneous:</u>
Skin ulcerations
Nasopharengeal distruction

<u>Visceral:</u>
Fever
Weight loss
Organ swelling
anaemia
Death

Challenges in *Leishmania* research

Difficult organism to study in the laboratory

- Slow growers
- Difficult to keep in culture
- Difficult to perform genetic studies
- Changes chromosomal ploidy when stressed

Overreaching goals:

Use genomic, proteomic and transcriptomic data to cluster genes according to protein expression pattern.

Example:

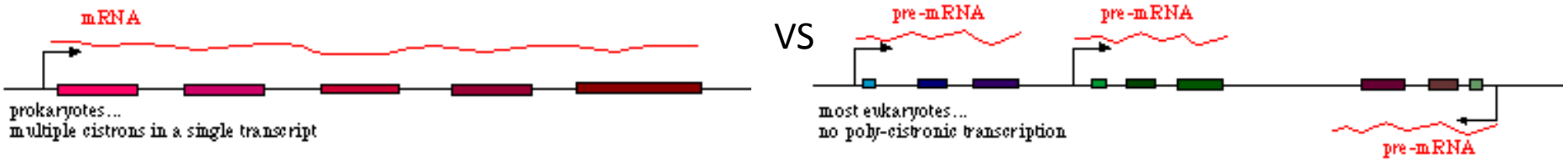
expressed in humans vs. expressed in sand flies

Verify gene assignment to a cluster by text mining publication records and assigning confidence of proper assignment to the individual genes.

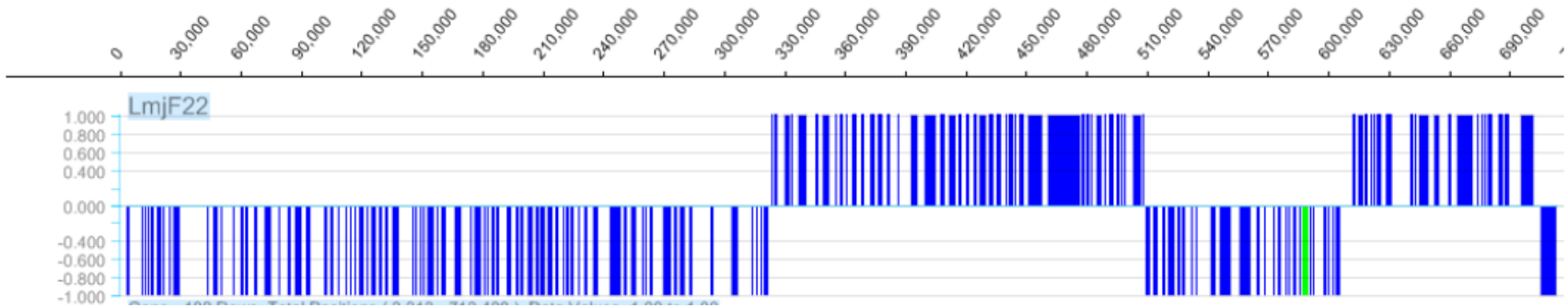
Identify patterns in genomic (DNA) transcriptomic (RNA) or proteomic (Protein) within gene groups that explain unique processes seen in trypanosomatids.

Unusual processes specific to Trypanosomatids

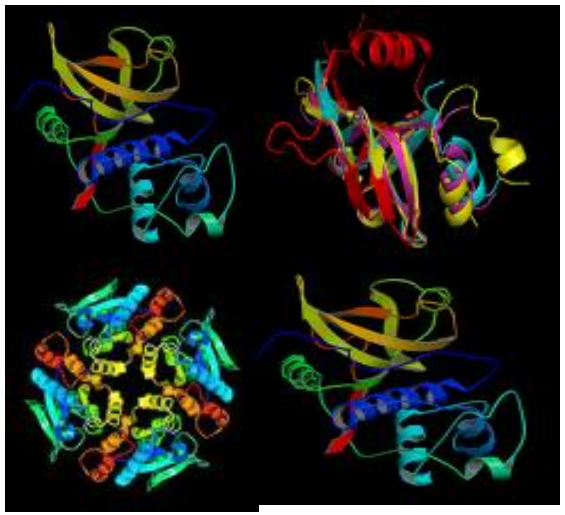
Polycistronic transcription

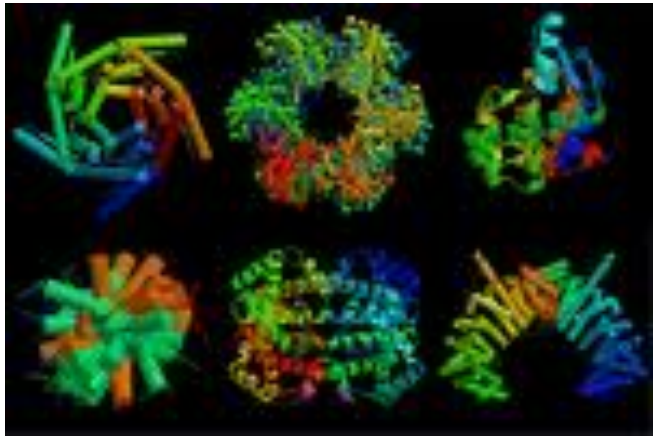


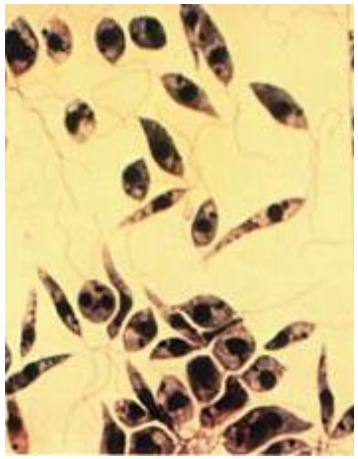
Only 2-3 promoters were identified



How do *Leishmania* change their protein profile between human and sandfly hosts?

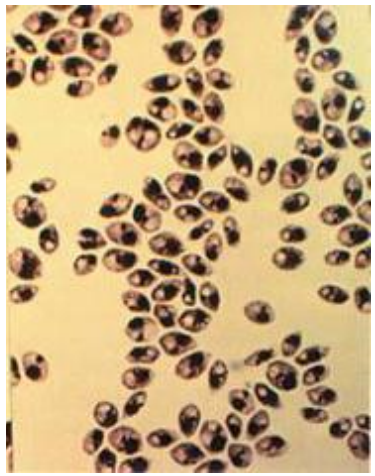






Promastigotes

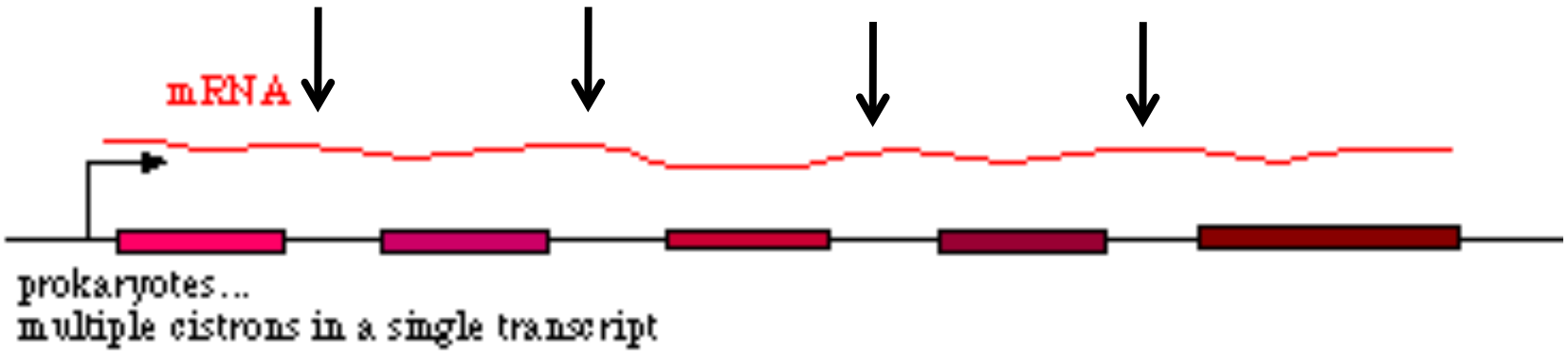
Amastigotes



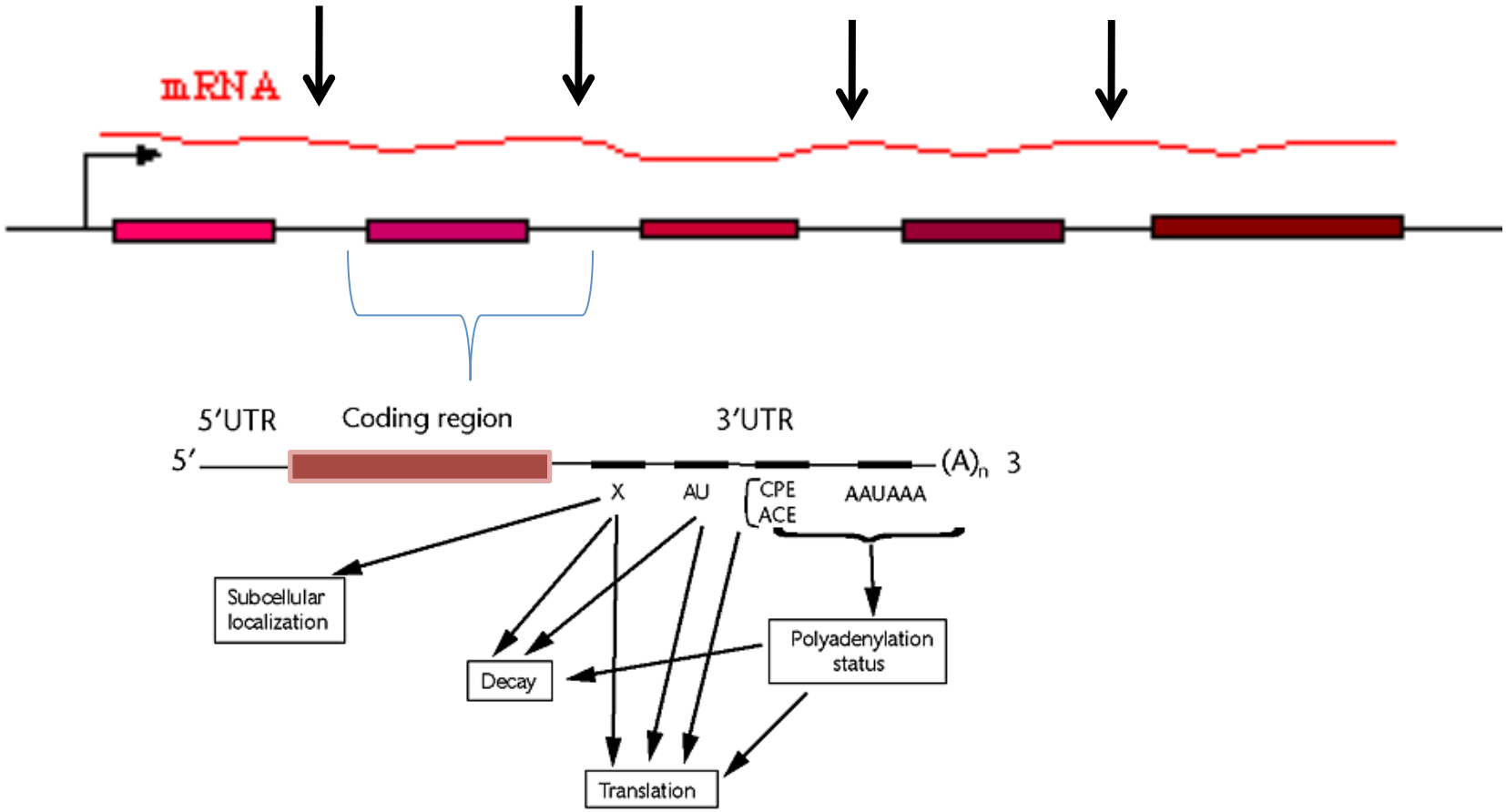
Approach: Identification of patterns or "motifs" within mRNA transcripts

Observation:

1. Steady levels of some mRNA throughout *Leishmania* lifecycles will explain expression of some of the genes.
2. For others, level of protein regulation must occur past the level of transcription.
3. Could regulation of protein expression reside in the Untranslated regions of the mRNA?



Approach: Identification of patterns or "motifs" within mRNA transcripts



Project Overview in Two Stages

Stage 1: Perform clustering and testing within Leishmania major Friedlin

- Assemble gene groups according to protein expression profiles (Sand fly vs. Human) and literature support.
- Define 3' UTRs, 5' UTRs, identify patterns between clusters and within subgroups.
- Test the effects of these patterns experimentally: i.e. siRNA knockdown for motifs that influence expression at differential stages.

Stage 2: Extend the process within and across trypanosomatids

- Use same cluster (or repeat clustering process depending upon organism)
- Define transcription stop sites, start sites (promoters)
- Look for patterns or motifs within proteins (differential degradation).
- Test experimentally

Part 1: Assemble expression groups

A. Make use of available genomic, proteomic and transcriptomic data to assemble initial groups ex. (human vs. sand fly)

B. Mine publication record for each individual gene:
    A. Rank confidence of assignment
    B. Form subgroupings within expression cluster

Part 1B mine publication records for each gene

- Previously: Targeted literature search for top 10 or 15 genes within each cluster.

- Currently: Literature search for each gene followed by ranking of information retrieved.
  - Pubmed searches with select MeSH terms to pull out initial pool of relevant abstracts.
  - Subsequent ranking of evidence level for each gene according to abstracts containing key
  - Include additional information such as GO terms to support assignments and identify potential false positives/ negatives.

Part 1B mine publication record for each individual gene

Pubmed search query using gene + X + Y + Z terms

Rank each article/abstract according to the presence of more detailed terms:

Examples:
Amastigote & expression & western
Amastigote
expression

Generate confidence values and sub groupings according to gene rankings.

Current progress

Stage 1: Perform clustering and testing within Leishmania major Friedlin

- Assemble gene groups according to protein expression profiles (Sand fly vs. Human) and literature support.
- Define 3' UTRs, 5' UTRs, identify patterns between clusters and within subgroups.
- Test the effects of these patterns experimentally: i.e. siRNA knockdown for motifs that influence expression at differential stages.

Stage 2: Extend the process within and across trypanosomatids

- Use same cluster (or repeat clustering process depending upon organism)
- Define transcription stop sites, start sites (promoters)
- Look for patterns or motifs within proteins (differential degradation).
- Test experimentally