Group project: Pitch

Context

You are a team of Bioinformaticians in a Biology lab.

The lab expertise is Biomedical microbiology.

Many researchers have increasingly access to genomics and omics data...
…But don't have the skills and time to learn a in-depth level of bioinformatics.

Your role is to support biologists in their research with data analyses and tools developments.

Two teams recently have had access to new sets of data, and need your help:

Project 1: Transcriptome analysis of *Trypanosoma congolense*

Project 2: Pathogenic Islands detection in a new strain of *Yersinia.*

Project 1: Transcriptome analysis of *Trypanosoma congolense*

Team A is working on Trypanosoma, a taxon of protists (Kinetoplastids, Euglenozoa, Discoba, Eukaryotes).
Trypanosomes are parasites, infecting different hosts and causing various diseases.
Ex: sleeping sickness, and Chagas disease.

The genome of these organisms are not easy and cost effective to sequences, but they need some genomic data.
The team is considering using ESTs.

EST: Expressed Sequences Tag:
short sub-sequence of a cDNA sequence
➔ *transcriptomics data*

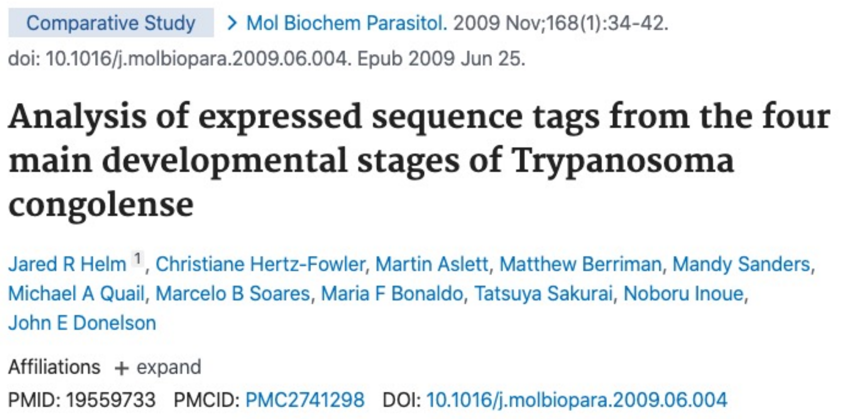
Project 1: Transcriptome analysis of *Trypanosoma congolense*

Team A is working on Trypanosoma, a taxon of protists (Kinetoplastids, Euglenozoa, Discoba, Eukaryotes).
Trypanosomes are parasites, infecting different hosts and causing various diseases.
Ex: sleeping sickness, and Chagas disease.

The genome of these organisms are not easy and cost effective to sequences, but they need some genomic data.
The team is considering using ESTs.

EST: Expressed Sequences Tag:
short sub-sequence of a cDNA sequence
➔ *transcriptomics data*

Helm et al. published in 2009 transcriptomic data for Trypanosoma congolese.



Comparative Study > Mol Biochem Parasitol. 2009 Nov;168(1):34-42.
doi: 10.1016/j.molbiopara.2009.06.004. Epub 2009 Jun 25.

Analysis of expressed sequence tags from the four main developmental stages of Trypanosoma congolense

Jared R Helm[1], Christiane Hertz-Fowler, Martin Aslett, Matthew Berriman, Mandy Sanders, Michael A Quail, Marcelo B Soares, Maria F Bonaldo, Tatsuya Sakurai, Noboru Inoue, John E Donelson

Affiliations + expand

PMID: 19559733 PMCID: PMC2741298 DOI: 10.1016/j.molbiopara.2009.06.004

Project 1: Transcriptome analysis of *Trypanosoma congolense*

Team A is working on Trypanosoma, a taxon of protists (Kinetoplastids, Euglenozoa, Discoba, Eukaryotes).
Trypanosomes are parasites, infecting different hosts and causing various diseases.

Ex: sleeping sickness, and Chagas disease.

The genome of these organisms are not easy and cost effective to sequences, but they need some genomic data.
The team is considering using ESTs.

EST: Expressed Sequences Tag:

short sub-sequence of a cDNA sequence

➔ *transcriptomics data*

Helm et al. published in 2009 transcriptomic data for Trypanosoma congolese.

The team wants to:

1/ Obtain the **proteins sequences from the ESTs data of T. congolese**
2/ Have a **pipeline of analysis of ESTs** data they can use for future datasets.
3/ For the protein sequences to be **named in a human-readable formatting**
4/ To be able to **retrieve a specific protein sequence** from the results.
5/ To **understand** the different outputs and how to use them.

Project 1: Transcriptome analysis of *Trypanosoma congolense*

The team wants to:

1/ Obtain the **proteins sequences from the ESTs data of T. congolese**
2/ Have a **pipeline of analysis of ESTs** data they can use for future datasets.
3/ For the protein sequences to be **named in a human-readable formatting**
4/ To be able to **retrieve a specific protein sequence** from the results.
5/ To **understand** the different outputs and how to use them.

You will:

1/ Dowload the raw ESTs data.
2/ Use *Transdecoder* to generate the CDS (coding protein sequences)
3/ Write a script to rename the sequences
4/ Write a script to retrieve sequences from a fasta file based on a ID number.
5/ Write a pipeline & manual to help the team repeat the analyses on a different dataset
6/ Write a report that explains your findings.

Project 1: Transcriptome analysis of *Trypanosoma congolense*

The team wants to:

1/ Obtain the **proteins sequences from the ESTs data of T. congolese**
2/ Have a **pipeline of analysis of ESTs** data they can use for future datasets.
3/ For the protein sequences to be **named in a human-readable formatting**
4/ To be able to **retrieve a specific protein sequence** from the results.
5/ To **understand** the different outputs and how to use them.

You will:

1/ Dowload the raw ESTs data.
2/ Use *Transdecoder* to generate the CDS (coding protein sequences)
3/ Write a script to rename the sequences
4/ Write a script to retrieve sequences from a fasta file based on a ID number.
5/ Write a pipeline & manual to help the team repeat the analyses on a different dataset
6/ Write a report that explains your findings.

To go above and beyond, you could:

- Do the same analyses for the 3 other ESTs datasets available in the paper
- Do an analysis of completeness of the transcriptome with *BUSCO*
- Prepare a pipeline to run the analyses on several datasets at once.