Biological Databases Final Project Proposal

Students:	Amy Li, Michelle, Ben Pote
Supervisor:	Ruslan Afasizhev
Contact:	617-414-1056
	ruslana@bu.edu

April 8, 2014

1 Project Overview

Trypansoma brucei (T. brucei) is a single cell protist parasite that gives rise to the African trypanosomiasis, or sleeping sickness. The mitochondrial genome of T. brucei consists of a network of circular structures called minicircles (1kb) and maxicircles (10kb). Maxicircles encode for protein-coding genes. However, such transcripts can be translated only after posttranscriptional deletion and insertion of uridines through a process called RNA-editing. This editing process is directed by short stretches of guideRNAs, which are encoded in minicircles. Understanding the organizational structures of minicircles and its encoded guideRNAs is an important step for deciphering the RNA-editing mechanism in T. brucei.

Here, we create a database for visualising next-generation sequencing data from mitochondrial minicircle DNA and small RNAs (putative guideRNAs) isolated from T. brucei. The database will facilitate the storage and analysis of minicircle sequences. Analyses to be performed include aligning reads and categorizing minicircles based on the minicircle conserved regions they contain.

Using MSAD, users would be able to retrieve and download information on minicircle sequences based on specific filtering criteria, including presence of conserved regions (CSBs), alignment coverage by smallRNA, etc. In addition, we wish to provide visualization for the minicircle sequence alignment information such as displaying the pileup of mapped sRNA reads along a specified minicircle query. If time permits, we will also visualize alignment for a given minicircle cluster.

With this database, users will be able to answer questions such as:

1. Which putative minicircle sequences contain conserved regions (CSB1, CSB2, CSB3)?

2. What is the multiple sequence alignment (MSA) for the specific cluster of sequences?

3. For a given minicircle sequence, which regions have high coverage of mapped sRNA reads? Where are the CSB regions (if any)?

Upon completion of the project, the database will be made accessible to the Afasizhev lab and BU affiliated students working on this dataset. If successful in this initial deployment, there is a possibility that the database could be made publicly available for other researchers to use.

Our database will contain data from the following:

1. PacBio Minicircle Sequencing Dataset - this dataset contains 39,939 filtered minicircle DNA sequences isolated from the mitochondria of Trypanosoma brucei. These reads are on average around 1kb long. For each sequence, we will have a unique sequence identifier, the actual DNA sequence, and a cluster assignment. The cluster assignment has already been precomputed using Connected Component Clustering, based on sequence similarity. We are currently in the process of applying other clustering methods, so we are interested in adding additional cluster assignments later on.

2. Illumina smallRNA-seq dataset - this contains 3,937,040 unique processed smallRNA reads which are putative guideRNAs. Each read has a duplication number (how many duplicate sequences of

itself is found in the original sequence file), an RNA sequence, and alignment information: where each sRNA read aligned to each of the minicircle sequences (some align to more than 1 or no minicircle sequences, and can align at more than two different positions within the same minicircle sequence). This is a very big file, so we may just store alignment information (processed from BED format) into the database and not each individual smallRNA sequence.

3. The sequences for the known minicircle conserved regions CSB1, CSB2, and CSB3. These are short sequences that are commonly found in curated minicircles (CSB3 is associated with the origin of replication). We'd like to be able to query which putative minicircle sequences contain each or combinations of these regions.

Multiple Sequence Alignments: given a cluster id, we can retrieve minicircles associated with such cluster and generate a Multiple Sequence Alignment of these sequences on the fly (or precomputed).

2   Tasks to accomplish

We plan to implement the following:

1. Create database tables (on bioed)

2. Ensure database conforms to ER diagram

3. Write cgi interface to database

   (a) About page (introduction)
   (b) A separate page for data retrieval/search, including:
       i. CSB based filtering
       ii. Coverage based filtering
       iii. Auto-complete features for searches
       iv. Dataset download ability
   (c) Multiple sequence alignment page:
       i. Display multiple sequence alignments
       ii. Download

4. Data download

5. Testing

3   ER diagram

Here we present an entity-relationship diagram that represents how we plan to build our database:
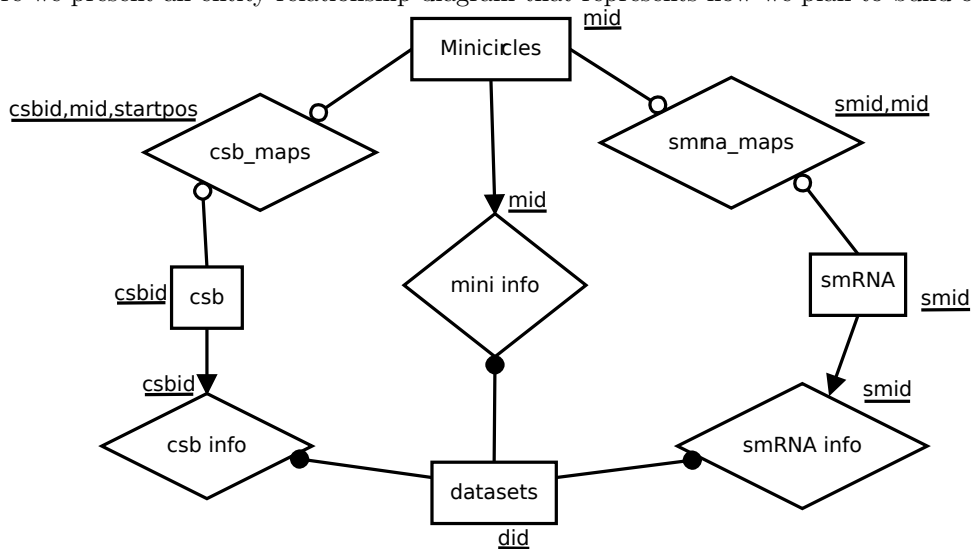
4 Tables

These are the tables we plan to implement in our project, with all fields listed. Primary keys are underlined, and foreign keys are italicized.

- **CSB:** <u>csbid</u>, **sequence**
- **smallRNA:** <u>smid</u>, **sequence, copynumber**
- **dataset:** <u>did</u>
- **minicircles:** <u>mid</u>, *did*, **sequence, clusternumber**
- **csb_maps:** *<u>csbid</u>,<u>mid</u>,<u>startpos</u>*, **endpos, strandinfo, quality**
- **smrna_maps:** *<u>smid</u>, <u>mid</u>*, **startpos, endpos, strandinfo, quality**

As indicated by the ER diagram above we will have tables for smRNAs, minicircles, CSBs, datasets, and the maps of both CSBs and smRNAs onto minicircles. These last two tables represent many-to-many relationships, and so they need their own table. The relationships between our three sequence tables (minicircles, smRNAs, and CSBs) and the dataset table are all one to many, as each sequence may only belong to one dataset, a dataset being here defined as describing the origin of the sequences, the platform they were sequenced on, and other similar metadata. This relationship will be contained in the table structure by having a dataset field holding a <u>did</u> in each of these sequence tables.