Introduction

Trypanosomes are parasitic protozoa that are transmitted by insects and infect many vertebrates with varying specificity across the species, causing Trypanosomiasis, also known as sleeping sickness in humans. Trypanosoma brucei brucei is one of the parasites infecting cattle causing a great economic burden in one of the poorer regions of the world. Of the two African cattle populations, N’dama and Zebu, only the Zebu cattle are susceptible to infection[1].

Methods

Data was acquired from the Sequence Read Archive as raw paired-end reads from Illumina HiSeq 1000, counting 42.8M reads. Tsetse flies (Glossina morsitans morsitans) were infected with T. b. brucei (RUMP 503 strain), and the salivary glands were dissected after 72h[3]. Data cleanup was performed by checking quality and locating adapters in FastQC and cutadapt was used to trim the adapters. Overlapping reads were aligned using flash and trimmed using prinseq

Mapping reads was done using tophat2 with T. brucei TREU927[2] as a template. Transcription level was estimated as FPKM (Fragments Per Kilobase of exon per Million fragments mapped) calculated by cufflinks. Non-translated RNA (rRNA, miRNA, etc.) and VSG genes were removed and peptide sequences were extracted using annotations. An FPKM cutoff was used to limit sequences to the top ~650 genes.

T cell epitopes were predicted using NetMHCIIpan for 15 of the 17 most frequent alleles described in Karimuribo et al. [4]. The allele names were translated from PCHR-RLFL numbers using the BoLA nomenclature on IPD-MHC (www.ebi.ac.uk/ipd/mhc). Two alleles were left out, because they were missing in IPD or NetMHCIIpan. These alleles were matched against the candidate transcripts as input. The output from NetMHCIIpan was processed by a simple implementation of the PopCover algorithm targeting only one genome. Epitopes were considered binders if their predicted affinity IC50 score was below 25 nM

B cell candidate epitopes were found by narrowing down the list of peptides using TMHMM to find genes containing possible transmembrane helices, and SignalP to locate possible signal peptides. Genes present in both SignalP and TMHMM output were deemed signal peptides and removed from the TMHMM output. Redundant sequences were removed from either list using skippredundant with an identity cutoff of 95%. Proteins present in organelle membranes, such as mitochondria or nucleus were removed from the TMHMM list using annotations. The possible membrane proteins were filtered one final time to extract proteins with a linear stretch of at least 20 residues on the outside of the outer membrane.

Results

During quality trimming 3.4% reads. 43.4% reads mapped to T. brucei. 8674 (97.7%) genes showed some coverage from cufflink mapping.

PopCover was run until every BoLADR3 allele was targeted at least twice, yielding 17 candidate peptides, coverage from cufflink mapping.

Discussion

The B cell candidates should be verified in the laboratory. Due to lack of PDB structures and an already limited number of proteins to investigate, we decided not to predict B cell epitopes to limit the candidates further. An alternative would be to use homology modeling, to predict the structures and then predict epitopes. Predicted surface proteins showed a significant redundancy, the source of which should be investigated, as it could be due to unannotated VSGs or a pipeline artefact.

There is some redundancy in the PopCover peptides, this could most likely have been avoided by more fine tuning of the algorithm. Due to way PopCover penalises selected alleles, it is not as effective at selecting the least selected allele after every allele has been selected at least once.

References