# Assessment of sequence descriptions of selected *Theileria parva* hypothetical proteins retrieved from sequence similarity search databases

**Mampa M. S[1], Mokoena. F[2], Matjila T. P[1], Sibeko K. P[1]**

[1]Department of Tropical Veterinary Diseases, University of Pretoria, Onderstepoort, South Africa, e-mail: mmogauselina@yahoo.com

[2]Department of Life and Consumer Science, University of South Africa, Johannesburg, South Africa.

## Introduction

*Theileria parva* is a tick-borne protozoan parasite transmitted by the brown ear tick, *Rhipicephalus appendiculatus.* The parasite is responsible for the cattle diseases, East Coast fever (ECF) and Corridor diseases, respectively caused by cattle- and buffalo-derived *T. parva* isolates [1]. The different disease outcomes resulting from *T. parva* infections led to the study of gene expression profiles during the schizont developmental stage (infective phase) of the parasite. Consequently, a transcriptome study was undertaken which identified 1089 differentially expressed genes (DEGs) between two *T. parva* isolates representing cattle- and buffalo-derived parasites. Analysis of DEGs showed that 74% (n=867) code for hypothetical proteins (HPs) (proteins with unknown functions), according to the published *T. parva* genome sequence [2]. These HPs could play a vital role in the pathogenicity and host-parasite interaction. In an attempt to functionally annotate selected *T. parva* HPs, it was discovered that most hits from outputs of automated sequence similarity search databases do not possess the acceptable sequence identity and coverage to the query. Thus, it became necessary to first assess sequence descriptions assigned to HPs by different databases in order to accurately predict possible biological roles of these proteins in *T. parva*.

## Study aim

To assess sequence descriptions (SDs) assigned to hypothetical proteins by automated sequence similarity search databases (ASSSDs).

## *In silico* approach

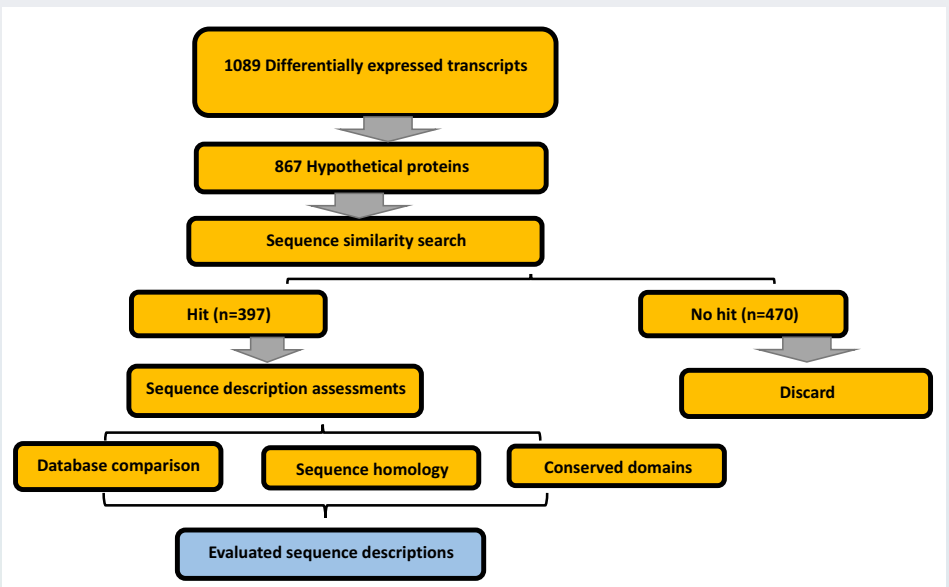


**Figure 1**: An outline of procedures followed to assess SDs assigned to hypothetical proteins, using *in silico* method

**Table 1:** List of bioinformatics tools and databases used for *in silico* analysis.

| Software/ Database | URL address | Annotation analysis type |
|---|---|---|
| **KEGG** | http://www.genome.jp/kegg/ | SDs & Pathways |
| **KOBAS** | http://kobas.cbi.pku.edu.cn | SDs & Pathways |
| **BLAST2G0** | https://www.blast2go.com | Gene ontology & SD |
| **BLAST** | https://blast.ncbi.nlm.nih.gov | Sequence homology |
| **Clustalw** | www.ch.embnet.org | Multiple sequence alignment |
| **Pfam** | http://pfam.sanger.ac.uk/ | Protein families and domains |
| **NCBI CD** | www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi | Conserved domains |
| **CDART** | http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi. | Domain architectures |
| **SMART** | http://smart.embl.de/ | Domain modular architectures |
| **INTERPRO** | http://www.ebi.ac.uk/interpro/about.html | Motifs and domains |

## Results and discussion

- Sequence similarity search using BLAST2GO, KEGG and KOBAS databases (ASSSDs) detected consensus SDs for 154 of the 397 *T. parva* HPs investigated **(Figure 2a)**.
- Inferring homology to related species confirmed SDs of 158 HPs **(Figure 2b)**, showing that 60% of SDs initially assigned by ASSSDs were not supported by the acceptable sequence homology criteria.
- *Theileria annulata* was the most useful homolog for inferring annotations; this parasite is the most closely related to *T. parva* and both have unique host cell transformation traits [2] **(Figure 3a)**.
- Conserved domain(s) are vital in discovering SDs as they are the elementary functional units of a protein [3]. As observed in this study, 91 of the 237 HPs that failed to meet the homology analysis criteria could be successfully assigned SDs from conserved domain analysis **(Figure 3b)**.
- Overall, 249 of 397 *T. parva* HPs analyzed in this study were successfully assigned SDs.

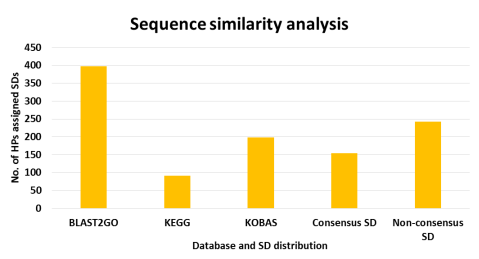


**Figure 2a:** Results obtained from sequence similarity search analysis using BLAST2G0, KOBAS and KEGG databases, showing the distribution of SDs assigned to selected *T. parva* HPs. SD= Sequence description;HPs= Hypothetical proteins.

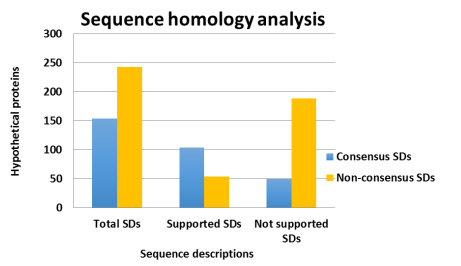


**Figure 2b**: The outcome of sequence homology analysis based on sequence identity of ⩾ 30% and ⩾ 50% coverage, showing the number of sequence descriptions supported by sequence homology analysis and non-supported sequence descriptions assigned to analyzed *T. parva* HPs.

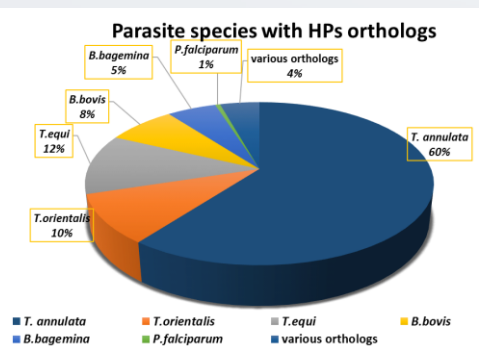


**Figure 3a:** Distribution of *T. parva* HPs (n=158) with orthologs in various related parasite species used to assign sequence description based on good homology (sequence identity ⩾ 30% and coverage ⩾50%).

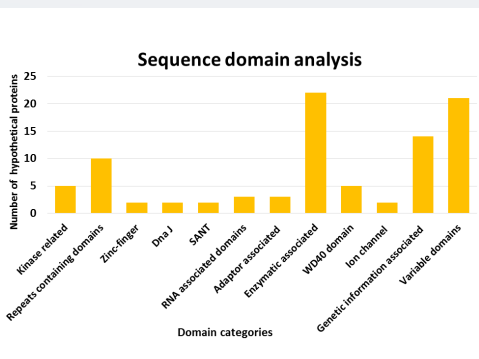


**Figure 3b**: A representation of different domains used to assign sequence descriptions of 91 HPs which failed to be annotated based on sequence homology criteria.

## Conclusion

The results from this study have showed that output from automated sequence similarity databases is not always reliable in assigning SDs for specific species, thus confirmation using other approaches is critical.

## Acknowledgments

We express our gratitude to the funding entities, NRF (Bioinformatics and Functional Genomics grant) and AgriSETA, as well as the University of Pretoria for providing the platform to accomplish the project.

## References

[1] Morzaria S.P., Katende J., Musoke A., et al., 1999, *Parasitologia* 41(1), 73-80.
[2] Gardner, M.J., Bishop, R., Shah, T., et al, 2005, *Science (New York, N.Y.)* 309(5731), 134-137.
[3] Marchler-Bauer, A., Zheng, C., Chitsaz, F., et al., 2012, *Nucleic acids research* gks1243.

photo: www.scpd.stanford.edu

National Research Foundation

AgriSETA

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA