Identification and annotation of alternative splicing of microexon genes in version 10 of the *Schistosoma mansoni* genome

Karolline Monteiro Vieira da Silva[1,2], Ana Carolina Tahira[1], Sergio Verjovski-Almeida[1,3]

[1]Laboratório de Ciclo Celular, Instituto Butantan, São Paulo 05503-900, SP, Brazil
[2]Programa Interunidades de Pós-Graduação em Bioinformática - USP, 05508-000, SP, Brazil
[3]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo 05508-000, SP, Brazil

Schistosomiasis is a disease caused by the trematode *Schistosoma mansoni*, which affects millions of people in tropical and subtropical regions and is considered a neglected tropical disease by the World Health Organization (WHO). The parasite can remain in the host for decades, evading the immune system. Currently, there are no vaccines available, and the only drug, praziquantel, is effective only against adult worms, without preventing reinfection. The efficiency of the immune evasion mechanism coordinated by the parasite suggests the possible use of alternatively-spliced microexon genes (MEGs), characterized by multiple exons of 6 to 36 nucleotides in tandem, which may be related to immune evasion mechanisms. This study aims to identify and annotate experimental evidence of alternative splicing of microexons in version 10 of the *S. mansoni* genome, using public RNA-Seq data. For this, a specific and sensitive pipeline was used to re-map RNA-Seq data to the genome (SM_V10, WormBase WBPS19) and identify microexons. The pipeline includes filtering and quality control (FastQC, fastp), followed by alignment with STAR and, in parallel, mapping with OLego. The results of both alignments are concatenated to create an annotated splicing index, which is applied in a second step with STAR, allowing the detection of both known and novel splicing sites. Parameterization tests were performed to optimize the sensitivity of the pipeline. In the first STAR pass, three groups of parameters were evaluated: “STANDARD” (least restrictive), ‘MICROEX’ (intermediate), and “MODEL” (most restrictive). At the same time, three variations of OLego were tested: with regression and annotation, with regression only, and without regression/annotation. In the second STAR pass, four groups were formed, each subdivided according to the indices. With the STANDARD parameters, an average of 204,215 junctions/sample was identified (among the three variations of the index), consistent with its lower restrictiveness and the presence of possible false positives. In the MICROEX I group, the average was 152,855 junctions/sample, and in the MICROEX II group, 155,218 junctions/sample. In the MODEL group, ~147,515 junctions/sample were detected. Transcript assembly was performed with StringTie, and the MICROEX configuration was chosen because it presented the best balance between sensitivity and efficiency. For OLego, processing using regression and annotation took 9 hours/sample, while using only regression or without regression/annotation took 2 hours/sample; thus, regression-only was chosen, considering that ~3,000 samples will be processed. Next, a filter was applied to identify MEGs, defining that: (i) maximum microexons size is 81 nt; (ii) transcripts with 3–4 internal exons must be formed only by microexons; (iii) transcripts with ≥5 internal exons must have the

largest block of tandem microexons representing at least 75% of the internal exons. Using these criteria, 70 transcripts associated with previously annotated MEGs were identified, of which 33% are new splice isoforms. The next steps include quantification with Ballgown and alternative splicing analysis with Whippet, focusing on MEGs. The work contributes to the understanding of possible molecular mechanisms of immune evasion by *S. mansoni* and points to possible new therapeutic targets.

Keywords: Microexons, MEGs, *Schistosoma mansoni*, Bioinformatics, RNA-seq, Alternative splicing

Funding: FAPESP grant 2018/23693-5; CNPq fellowship to KMVS.

References

Brasil, M. da S. (2022). *Boletim Epidemiológico - Situação epidemiológica da esquistossomose mansoni no Brasil, 2010 a 2022*. https://www.gov.br/saude/pt-br/assuntos/boletins-epidemiologicos

Brasil, M. da S. (2024). *Diretrizes técnicas -vigilância da esquistossomose mansoni*. https://bvsms.saude.gov.br/bvs/publicacoes/vigilancia_esquistossome_mansoni_diretrizes_tecnicas.pdf_1ed.pdf

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Liu, Z., Zhu, C., Steinmetz, L. M., & Wei, W. (2023). Identification and quantification of small exon-containing isoforms in long-read RNA sequencing data. *Nucleic Acids Research*, *51*(20), E104. https://doi.org/10.1093/nar/gkad810

Smith, A. D., & de Sena Brandine, G. (2021). Falco: High-speed FastQC emulation for quality control of sequencing data. *F1000Research*, 8, 1874. https://doi.org/10.12688/f1000research.21142.2

Sterne-Weiler, T., Weatheritt, R. J., Best, A. J., Ha, K. C. H., & Blencowe, B. J. (2018). Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Molecular Cell*, *72*(1), 187-200.e6. https://doi.org/10.1016/j.molcel.2018.08.018

WHO. (2022). *WHO guideline on control and elimination of human schistosomiasis*. https://www.who.int/publications/i/item/9789240041608

WHO. (2023, February 1). *Schistosomiasis*. https://www.who.int/News-Room/Fact-Sheets/Detail/Schistosomiasis.

Wu, J., Anczuków, O., Krainer, A. R., Zhang, M. Q., & Zhang, C. (2013). OLego: Fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Research, 41(10), 5149–5163. https://doi.org/10.1093/nar/gkt216

Yu, H., Li, M., Sandhu, J., Sun, G., Schnable, J. C., Walia, H., Xie, W., Yu, B., Mower, J. P., & Zhang, C. (2022). Pervasive misannotation of microexons that are evolutionarily conserved and crucial for gene function in plants. *Nature Communications*, 13(1), 820. https://doi.org/10.1038/s41467-022-28449-8