

FLUXO COMPUTACIONAL SISTEMATIZADO PARA PRÉ-PROCESSAMENTO E VALIDAÇÃO DA QUALIDADE DE DADOS TRANSCRIPTÔMICOS DE *TOXOPLASMA gondii*

Luís Henrique Féres Junqueira Filho¹.

¹Universidade Federal de Uberlândia (UFU), Uberlândia, MG. <http://lattes.cnpq.br/6839323444637486>

PALAVRAS-CHAVE: Bioinformática. Transcriptômica. Toxoplasmose.

DOI: 10.47094/IVCOLUBRASC.2026/RE/2

INTRODUÇÃO

O *Toxoplasma gondii* (*T. gondii*), organismo causador da toxoplasmose, é um parasito intracelular obrigatório, capaz de infectar virtualmente todos os vertebrados homeotérmicos, incluindo os seres humanos. Os felinos, especialmente os gatos domésticos, constituem seus hospedeiros definitivos, nos quais ocorre a reprodução sexuada do parasito, enquanto os demais animais atuam como hospedeiros intermediários. A ampla prevalência global da doença, aproximadamente 30% da população mundial (BIGNA et al., 2020), e a capacidade do parasito de infectar diferentes espécies conferem à toxoplasmose elevada relevância epidemiológica, constituindo uma importante questão de saúde pública em diferentes regiões do mundo (HATAMNAHAVANDI et al., 2021).

A infecção por *T. gondii* apresenta manifestações clínicas variadas e pode ocorrer de forma assintomática ou com sintomas inespecíficos. Determinados grupos populacionais apresentam maior risco de desenvolver formas graves da doença. Em indivíduos imunocomprometidos, como pessoas submetidas a transplantes, pacientes em tratamento imunossupressor e indivíduos vivendo com condições que comprometem a resposta imunológica, a reativação ou a infecção pelo parasito pode resultar em manifestações severas, incluindo comprometimento do sistema nervoso central e outros órgãos. Uma infecção adquirida durante a gestação apresenta especial importância devido ao risco de transmissão vertical, podendo ocasionar toxoplasmose congênita e consequências que incluem aborto, alterações neurológicas, comprometimento ocular e outras manifestações que podem persistir ou surgir ao longo do desenvolvimento da criança (WEISS; DUBEY, 2009).

Nesse contexto, a compreensão dos mecanismos biológicos e moleculares envolvidos na infecção por *T. gondii* é fundamental para ampliar o conhecimento acerca da patogênese da toxoplasmose e contribuir para o desenvolvimento de estratégias de prevenção, diagnóstico e tratamento. O aumento de dados provenientes de estudos de sequenciamento genético e da disponibilidade de bancos de dados públicos têm expandido as possibilidades de investigação dos processos biológicos associados ao parasito.

A bioinformática constitui uma importante estratégia para a análise *in silico* de dados biológicos, permitindo o processamento, a organização e a investigação de grandes volumes

de informações provenientes de estudos genômicos e transcriptômicos, fornecendo subsídios para a identificação de genes e mecanismos potencialmente associados à patogênese (HASSAN et al., 2012). A utilização de abordagens computacionais possibilita explorar mecanismos moleculares relacionados à interação parasito-hospedeiro, identificação de genes e vias potencialmente associadas à patogênese e resposta à infecção, além de contribuir para a identificação de possíveis alvos terapêuticos. A análise transcriptômica de *T. gondii* constitui uma importante estratégia para compreender os mecanismos moleculares associados ao desenvolvimento, diferenciação e interação parasito-hospedeiro. Estudos baseados em RNA-seq demonstram que alterações na expressão gênica acompanham diferentes estágios do ciclo de vida do parasito, contribuindo para a compreensão de processos relacionados à infecção e persistência no hospedeiro (RADKE et al., 2005).

Assim, o estabelecimento de fluxos computacionais padronizados e reprodutíveis para o processamento de dados transcriptômicos de *T. gondii* torna-se uma etapa fundamental para a obtenção de dados confiáveis e adequados às análises subsequentes. Neste trabalho explorou-se o desenvolvimento de uma abordagem padronizada para o pré-processamento de dados transcriptômicos de *T. gondii*, abrangendo desde a coleta de dados públicos, a conversão dos arquivos de sequenciamento, avaliação e controle de qualidade dos dados obtidos, remoção de sequências de baixa qualidade e posterior alinhamento ao genoma de referência, estabelecendo uma base sólida para a realização de análises transcriptômicas posteriores que possibilitem a sua aplicação na investigação dos potenciais mecanismos moleculares relacionados à toxoplasmose.

OBJETIVO

Descrever as etapas necessárias de pré-processamento de dados transcriptômicos de *T. gondii*, utilizando-se de ferramentas de bioinformática, que possibilitem a realização de análises posteriores com o emprego de dados com a qualidade adequada.

METODOLOGIA

Os dados de sequenciamento destinados ao pré-processamento foram obtidos a partir do NCBI (*National Center for Biotechnology Information*), por meio do repositório *Sequence Read Archive* (SRA), sendo selecionados conjuntos de dados provenientes de experimentos de RNA-seq disponíveis publicamente. O pré-processamento de dados transcriptômicos foi estabelecido pelas seguintes etapas: (I) Adequação ao formato de dados FASTQ, (II) Avaliação da qualidade dos dados, (III) Controle de qualidade das bases nitrogenadas, e (IV) Alinhamento de sequências.

RESULTADOS E DISCUSSÃO

Etapa (I): Os arquivos no formato SRA foram inicialmente coletados e processados utilizando o SRA Toolkit (SAYERS et al., 2022), ferramenta que possibilita a conversão dos arquivos para o formato FASTQ, que é o padrão adequado às etapas subsequentes de

análise.

Etapa (II): Após a conversão, os arquivos FASTQ foram submetidos à avaliação de qualidade utilizando o software FastQC (DE SENA BRANDINE et al., 2021), permitindo a inspeção de parâmetros relacionados à qualidade das sequências, conteúdo de bases nitrogenadas, distribuição de qualidade por posição, presença de adaptadores e outras características relevantes dos dados brutos. Os resultados dessa etapa foram utilizados para identificar possíveis problemas de qualidade dos dados obtidos anteriormente.

Etapa (III): A seguir, realizou-se o controle de qualidade das bases nitrogenadas utilizando o software Trimmomatic (BOLGER et al., 2014), com o objetivo de remover as sequências de adaptadores e bases nitrogenadas de baixa qualidade, além de reduzir a presença de regiões inadequadas para as análises posteriores. Ao final dessa etapa, foram obtidos arquivos contendo as bases nitrogenadas filtradas e de maior qualidade, os quais foram utilizados como entrada para o alinhamento.

Etapa (IV): Finalizando, o alinhamento das sequências filtradas foi realizado utilizando o software STAR (*Spliced Transcripts Alignment to a Reference*), ferramenta desenvolvida para o alinhamento de dados de RNA-seq contra um genoma de referência (DOBIN et al., 2012). Para esta etapa, foi utilizado o genoma de referência de *T. gondii*, permitindo o mapeamento das bases nitrogenadas e a geração dos arquivos de alinhamento necessários às análises transcriptômicas subsequentes.

Como resultado, estruturou-se o passo a passo dos procedimentos em etapas sequenciais constituído pela coleta dos dados no NCBI/SRA, conversão para FASTQ, avaliação da qualidade, filtragem e remoção de sequências de baixa qualidade e alinhamento ao genoma de referência, estabelecendo um fluxo de processamento padronizado e reproduzível para a preparação dos dados transcriptômicos de *T. gondii*.

CONSIDERAÇÕES FINAIS

Ferramentas de bioinformática permitem o processamento e a análise de grandes volumes de dados de sequenciamento, possibilitando sua utilização em investigações *in silico*. Porém, o pré-processamento adequado desses dados torna-se fundamental, uma vez que etapas de controle de qualidade, remoção de sequências inadequadas e alinhamento contribuem para a obtenção de dados confiáveis para a realização das análises em etapas posteriores.

O desenvolvimento sequencial proposto estabeleceu um fluxo sistematizado para o pré-processamento de dados transcriptômicos de *T. gondii*, organismo em análise, desde a coleta dos dados públicos, até seu alinhamento ao genoma de referência. A utilização integrada das ferramentas de bioinformática possibilitou organizar e qualificar os dados para serem utilizados em etapas analíticas subsequentes, contribuindo para estudos futuros de entendimento dos mecanismos moleculares envolvidos na toxoplasmose. Considera-se que a abordagem adotada poderá constituir uma etapa inicial para as investigações voltadas à identificação de potenciais alvos terapêuticos, biomarcadores e mecanismos relacionados

à infecção por *T. gondii*, reforçando a aplicação da bioinformática como ferramenta de apoio à pesquisa em saúde coletiva e combate às infestações parasitárias.

PRINCIPAIS REFERÊNCIAS

- BIGNA, J. J. *et al.* Global, regional, and country seroprevalence of *Toxoplasma gondii* in pregnant women: a systematic review, modelling and meta-analysis. **Scientific Reports**, v. 10, n. 1, 21 jul. 2020. DOI: <https://doi.org/10.1038/s41598-020-69078-9>
- BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 1 abr. 2014. DOI: <https://doi.org/10.1093/bioinformatics/btu170>
- DE SENA BRANDINE, Guilherme; SMITH, Andrew D. Falco: high-speed FastQC emulation for quality control of sequencing data. **F1000Research**, v. 8, p. 1874, 27 jan. 2021. DOI: <https://doi.org/10.12688/f1000research.21142.2>
- DOBIN, Alexander *et al.* STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, v. 29, n. 1, p. 15–21, 25 out. 2012. DOI: <https://doi.org/10.1093/bioinformatics/bts635>
- HASSAN, M. A. *et al.* De novo reconstruction of the *Toxoplasma gondii* transcriptome improves on the current genome annotation and reveals alternatively spliced transcripts and putative long non-coding RNAs. **BMC genomics**, 12 dez. 2012. DOI: <https://doi.org/10.1186/1471-2164-13-696>
- HATAM-NAHAVANDI, K. *et al.* *Toxoplasma gondii* infection in domestic and wild felids as public health concerns: a systematic review and meta-analysis. **Scientific Reports**, v. 11, n. 1, 4 maio 2021. DOI: <https://doi.org/10.1038/s41598-021-89031-8>
- RADKE, J. R. *et al.* The transcriptome of *Toxoplasma gondii*. **BMC Biology**, v. 3, n. 1, p. 26, 2 dez. 2005. DOI: <https://doi.org/10.1186/1741-7007-3-26>
- SAYERS, Eric W.; O’SULLIVAN, Chris; KARSCH-MIZRACHI, Ilene. Using GenBank and SRA. **Methods in Molecular Biology**, p. 1–25, 2022. DOI: https://doi.org/10.1007/978-1-0716-2067-0_1
- WEISS, L. M.; DUBEY, J. P. Toxoplasmosis: a history of clinical observations. **International journal for parasitology**, v. 39, n. 8, p. 895–901, 1 jul. 2009. DOI: <https://doi.org/10.1016/j.ijpara.2009.02.004>