

ANÁLISE DE PARÂMETROS DE FERTILIDADE EM *Equus caballus* UTILIZANDO MODELOS NÃO SUPERVISIONADOS DE *MACHINE LEARNING*

LUANA CARLA SALVI¹; LUCAS PETITEMBERTE DE SOUZA²; WILLIAM BORGES DOMINGUES³; IZANI BONEL ACOSTA⁴; LUIZ ALEXANDRE CHISINI⁵; VINICIUS FARIAS CAMPOS⁶

¹Universidade Federal de Pelotas – luanacarlalvi@gmail.com

²Universidade Federal de Pelotas – lucasouza.contato@gmail.com

³Universidade Federal de Pelotas – williamwwe@gmail.com

⁴Universidade Federal de Pelotas – izanibonel@hotmail.com

⁵Universidade Federal de Pelotas – alexandrechisini@gmail.com

⁶Universidade Federal de Pelotas – fariascampos@gmail.com

1. INTRODUÇÃO

O setor equestre brasileiro movimenta anualmente mais de R\$16 bilhões e é responsável pela geração de cerca de 3 milhões de empregos (MAPA, 2016). Ao contrário de outras espécies de interesse comercial, onde a seleção de machos reprodutores se baseia na fertilidade, a escolha dos garanhões é determinada principalmente pelo pedigree, características fenotípicas e desempenho atlético (VARNER et al., 2015).

A análise espermática em garanhões geralmente inclui a avaliação da motilidade total e progressiva, concentração de espermatozoides, morfologia, além de parâmetros cinéticos como velocidade média, velocidade em linha reta, velocidade curvilínea, retidão e linearidade (MEDICA et al., 2023). No entanto, as análises convencionais, tanto macroscópicas quanto microscópicas, têm mostrado limitações na predição precisa da fertilidade (SALAS-HUETOS et al., 2023). Os parâmetros usados nessas análises não possuem a sensibilidade diagnóstica necessária para diferenciar diferentes graus de fertilidade, uma demanda comum na reprodução equina (VARNER et al., 2015).

Nas últimas décadas, houve um aumento significativo no uso de biotecnologias reprodutivas no setor equestre, impulsionado pelo alto valor econômico desses animais e pela demanda crescente por cavalos de elite com excelente desempenho (AZCONA et al., 2020). Nesse cenário, a identificação de biomarcadores de fertilidade, como os miRNAs, têm o potencial de melhorar a avaliação reprodutiva dos garanhões, visto que os métodos tradicionais apresentam baixa sensibilidade, o que pode acarretar prejuízos econômicos ao setor.

O trabalho teve o objetivo utilizar um método não supervisionados de *Machine Learning* para identificar amostras de alta e baixa fertilidade em garanhões, com base no conjunto de dados fornecido pelas análises de motilidade e cinética espermática, fornecidas pelo sistema CASA (*Computer-Assisted Sperm Analysis*). Este estudo foi um projeto piloto do experimento que visa identificar miRNAs espermáticos biomarcadores para análise de fertilidade em garanhões utilizando qPCR.

2. METODOLOGIA

Foram analisadas amostras seminais de 28 garanhões da raça Crioula, clinicamente saudáveis e sexualmente ativos, fornecidas pela empresa Hartwig

Fertilidade Equina, situada em Pelotas-RS. Um ejaculado foi obtido de cada garanhão utilizando vagina artificial, imediatamente diluído, na proporção de 1:1 (diluyente: sêmen) utilizando diluyente comercial Botu-Sêmen® (Botupharma, Botucatu, SP, Brasil) e acondicionado em recipiente comercial para transporte das amostras a 15°C até o laboratório.

Os parâmetros de motilidade e cinética espermática foram analisados utilizando metodologia descrita por Domingues e colaboradores (2021). As amostras foram avaliadas utilizando o sistema computadorizado CASA (SpermVision®, Minitube, Alemanha). Foram avaliados os seguintes parâmetros: motilidade total (TM) (%) e progressiva (PM) (%), distância média percorrida (DAP) (µm), distância curvilínea (DCL) (µm), distância retilínea (DSL) (µm), velocidade média percorrida (VAP) (µm/s), velocidade curvilinear (VCL) (µm/s), velocidade linear (VSL) (µm/s), taxa de retilinearidade (STR) (%), linearidade (LIN) (%), coeficiente de oscilação (WOB) (%), amplitude do deslocamento lateral da cabeça (ALH) (µm) e frequência do batimento flagelar (BCF) (Hz).

A análise de dados foi realizada em Python, onde os dados foram inicialmente pré-processados. As variáveis contínuas foram padronizadas utilizando o Z-Score, com o objetivo de evitar que variáveis em diferentes escalas influenciassem indevidamente os resultados dos agrupamentos. Inicialmente, aplicou-se uma análise de agrupamento hierárquico, através do método aglomerativo, utilizando um dendrograma para visualizar a estrutura hierárquica, ilustrando a relação entre as amostras e a distância de agrupamento. A análise não hierárquica foi realizada utilizando o algoritmo K-means da biblioteca Scikit-learn. Ambos os métodos utilizaram os dados padronizados. Posteriormente, o método do Elbow foi empregado para confirmar o número ótimo de clusters na estrutura não hierárquica. Para avaliar a qualidade dos clusters gerados, um gráfico da silhueta foi construído, permitindo visualizar a separação e a compactação dos grupos. O coeficiente de silhueta foi calculado para quantificar a qualidade dos agrupamentos, fornecendo uma medida da separação entre os clusters. Para verificar se as variáveis analisadas diferiram significativamente entre os clusters, aplicou-se uma ANOVA para cada uma das variáveis selecionadas. O agrupamento resultante foi analisado de forma bivariada em relação a cada característica associada à motilidade celular e a outros parâmetros de cinética espermática. Para cada variável, a estatística F e o valor-p foram calculados, permitindo identificar aquelas que mais contribuíram para a separação dos grupos e destacando as de maior poder discriminante entre os clusters.

3. RESULTADOS E DISCUSSÃO

A análise visual do dendrograma indicou que o agrupamento em três clusters foi a melhor forma de representar a distribuição dos dados. A análise de agrupamento utilizando o método K-means confirmou três clusters como o número ideal, conforme sugerido pela curva de Elbow. Este método considera a soma das distâncias internas (WCSS) entre as observações e seus respectivos centroides, e busca o ponto de inflexão na curva, que indica o número adequado de clusters. O método Elbow mostrou uma queda acentuada no WCSS até três clusters, sugerindo que esse número captura adequadamente as variações nos dados.

Para avaliar a qualidade dos agrupamentos gerados, foi utilizado o coeficiente de silhueta, que fornece uma métrica de separação entre os clusters,

juntamente com a construção do gráfico da silhueta para visualizar a compactação e a separação entre os grupos. Os resultados do coeficiente de silhueta foram os seguintes: 2 clusters: 0,4395; 3 clusters: 0,3919. Embora o valor do coeficiente de silhueta para 2 clusters tenha sido ligeiramente maior (0,4395), indicando uma melhor separação e compactação geral dos grupos, optou-se pela divisão em 3 clusters por diversos motivos práticos e analíticos. Primeiramente, o método do Elbow e a análise hierárquica através do dendrograma indicaram que a estrutura dos dados era mais bem representada com três grupos. A escolha por 3 clusters permite uma segmentação mais rica e biologicamente relevante dos dados, fornecendo uma discriminação mais detalhada entre diferentes níveis de qualidade espermática. Com 3 clusters, foi possível identificar padrões claros.

O primeiro grupo, denominado "alta fertilidade", inclui 15 amostras caracterizadas por valores elevados em diversas variáveis relacionadas à motilidade, como TM (motilidade total), PM (motilidade progressiva), VSL (velocidade em linha reta) e VAP (velocidade média do trajeto). Este cluster representa indivíduos com alta qualidade espermática, evidenciada pela maior velocidade e motilidade progressiva. O segundo grupo, identificado como "fertilidade intermediária", contém 8 amostras. Ele se diferencia por apresentar valores intermediários de TM e PM, além de características moderadas em termos de DAP (deslocamento curvilíneo médio) e DSL (deslocamento retilíneo). Essas amostras exibem velocidades mais baixas, como VAP e VCL (velocidade curvilínea), quando comparadas ao cluster de alta fertilidade, sugerindo uma qualidade espermática moderada. O terceiro grupo, denominado "baixa fertilidade", é composto por 5 amostras, caracterizado pelos menores valores de TM e PM, além de apresentar as menores velocidades (VAP e VCL) e parâmetros reduzidos, como DAP e DSL. Esse cluster reflete amostras com qualidade espermática significativamente inferior.

Com base nesses clusters, aplicou-se a ANOVA unidirecional para verificar quais variáveis se diferenciavam significativamente entre os grupos. Os resultados da ANOVA indicaram que variáveis relacionadas à cinética espermática apresentaram diferenças significativas entre os clusters. As variáveis com maior poder discriminante incluíram a VCL ($F = 52.86$, $p < 0.0001$), DCL ($F = 52.23$, $p < 0.0001$) e DAP ($F = 49.00$, $p < 0.0001$), seguidas por outros parâmetros de cinética e motilidade espermática. A motilidade é um dos principais parâmetros utilizados para categorizar amostras de alta e baixa fertilidade (TURRI et al., 2021). No entanto, apenas a motilidade total foi significativa para a clusterização das amostras ($p = 0.0001$), a motilidade progressiva (PM) ($p = 0.056$) e WOB ($p = 0.152$) não foram significativos para a diferenciação dos grupos em nossa análise. Nossos resultados corroboram parcialmente os achados de Trinquê e colaboradores (2023) que observaram correlação positiva entre TM, PM e espermatozoides rápidos (RAP) com a fertilidade do sêmen do garanhão, sendo a variável RAP (que considerou $VAP > 70 \mu\text{m/s}$ e $STR > 80\%$) a mais consistente com a fertilidade.

Os resultados confirmam que as variáveis de velocidade curvilínea (VCL), distância curvilínea (DCL) e distância média percorrida (DAP) desempenham papéis críticos na separação dos grupos, sugerindo que diferentes perfis de cinética celular estão associados a diferentes condições de fertilidade. Variáveis como VAP, VSL, DSL, ALH, BCF, LIN, STR e a tradicional motilidade total (TM) também apresentaram diferenças significativas entre os grupos, reforçando a

importância da análise integrada de parâmetros para a categorização das amostras.

A separação clara entre os clusters com base nas variáveis avaliadas destaca o potencial do uso de algoritmos não supervisionados para identificar padrões biológicos complexos. A análise de variância revelou um grupo de variáveis particularmente sensíveis para a diferenciação dos perfis celulares, o que pode ser útil para investigações futuras.

4. CONCLUSÕES

Esses resultados destacam a importância de múltiplos parâmetros no agrupamento de células em diferentes perfis de fertilidade e sugerem que variáveis relacionadas à cinética espermática são os principais discriminantes entre os grupos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AZCONA, F. V., M.; MOLINA, A.; TRIGO, P.; PERAL-GARCÍA, P. et al. Impact of reproductive biotechnologies on genetic variability of Argentine Polo horses. **Livestock Science**, vol. 231, n. 103848, 2020.

DOMINGUES, W. B.; SILVEIRA, T. L. R.; NUNES, L. S.; BLODORN, E. B. et al. GH Overexpression Alters Spermatic Cells MicroRNAome Profile in Transgenic Zebrafish. **Front Genet**, 12, p. 704778, 2021.

MAPA - MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO. **Revisão do estudo do complexo do agronegócio do cavalo**. Brasília, DF: MAPA, 2016.

MEDICA, A. J.; LAMBOURNE, S.; AITKEN, R. J. Predicting the Outcome of Equine Artificial Inseminations Using Chilled Semen. **Animals** (Basel), vol. 13, n. 7, 2023.

SALAS-HUETOS, A.; RIBAS-MAYNOU, J.; MATEO-OTERO, Y.; TAMARGO, C. et al. Expression of miR-138 in cryopreserved bovine sperm is related to their fertility potential. **J Anim Sci Biotechnol**, vol. 14, n. 1, p. 129, 2023.

TRINQUE, C. M. ; SILVA, L. S. ; PEREIRA, R. R. ; CANUTO, L. E. F. ; CAVALERO, T. ; FREITAS, C. P. ; PAPA, F. O. . Correlation of equine sperm kinetic parameters with fertility index. **Journal of Equine Veterinary Science**, v. 125, p. 46, 2023.

TURRI, F; CAPRA, E; LAZZARI, B; et al. A combined flow cytometric semen analysis and miRNA profiling as a tool to discriminate between high-and low-fertility bulls. **Frontiers in Veterinary Science**. v.8, 1–12, 2021.

VARNER, D. D.; GIBB, Z.; AITKEN, R. J. Stallion fertility: a focus on the spermatozoon. **Equine Vet J**, vol. 47, n. 1, p. 16-24, 2015.