Appendix: Downstream Analysis of DeepTAPE

Tongfei Shen[1+], Miaozhe Huo[1+], Wan Nie[1], Kaiqi Li[1], Ziwei Ma[1],Xikang Feng[2*], Shuai Cheng Li[1*]
[1]Department of Computer Science, City University of Hong Kong, Kowloon, China
[2]School of Software, Northwestern Polytechnical University, Xi' an, Shaanxi, China
https://github.com/deepomicslab/DeepTAPE-1.0

I. Results

A. The role of essential CDR3 3-mer oligopeptides in enhancing diagnostic discrimination

In the third fold of the DeepTAPE test set, each SLE patient sample's highest-scoring 2,000 sequences were selected, mask each 3-mer oligopeptide, and reinserted into the model for prediction. This process yielded high-scoring (0.7) and high-frequency (>40) 3-mer oligopeptides (Fig. 1A), which exhibit significant frequency differences between SLE patients and HI, influencing the overall sample score.

A set of essential 3-mer oligopeptides met the criteria: A-F-F, L-F-F, I-Y-F, Y-T-F. The proportion of sequences containing these 3-mer oligopeptides in SLE samples is significantly higher than in HI (p <0.001), and the total frequency of sequences containing the 3-mer oligopeptides is also significantly higher in SLE samples (p <0.001) (Fig. 1B). Focusing on the diagnostic discrimination ability of these 3-mer oligopeptides, classification based on the frequency of sequences containing the 3-mer oligopeptides in SLE patients and HI samples resulted in an AUC of over 63% for all 3-mer oligopeptides. Notably, the classification result for L-F-F achieved an impressive AUC of 81.9% , and the total frequency of the 3-mer oligopeptides pushed the AUC to 84.0% , making it a promising diagnostic feature (Fig. 1C).

B. Potential antigens and genes for SLE unearthed from significant sequences identified by deep learning

Filtering out the highest-scoring 2,000 sequences from the DeepTAPE test set of the third fold for each SLE patient sample, we obtained a collection of TCR sequences with high probability of contributing to SLE positivity. It is plausible that some of these sequences may interact with epitopes of SLE antigens, leading to immune system attacks on self-tissues or excessive TCR reactions to certain pathogenic antigenic, potentially resulting in excessive immune responses and the pathogenesis of SLE.

Based on these significant TCR sequences, a qualitative analysis is conducted through TCRanno's tcr2tcr function to obtain epitopes, antigens, and organism that perfectly match the sequences [1]. A filter condition is set to retain only those where the organism is Homo Sapiens.

Subsequently, by querying the GeneCards database, we identified the genes corresponding to the antigens and their associated diseases [2]. Through meticulous scrutiny of academic literature, we filtered out antigens that are relevant to autoimmune disorders (TABLE I). These antigens represent potential candidates for future clinical and experimental validation as potential targets in SLE [3, 4, 5, 6, 7, 8, 9, 10].

II. Discussion and Conclusion

The MASK method identified four essential 3-mer oligopeptides (A-F-F, L-F-F, I-Y-F, Y-T-F) characteristic of SLE patients. These oligopeptides' frequency can achieve an AUC of 84.0% , serving as a preliminary SLE screening classifier. Further, high-scoring sequences predicted by DeepTAPE, matched by TCRanno, and screened by GeneCards identified potential pathogenic genes and antigens for SLE, aiding in exploring its complex pathogenesis.

III. Acknowledgements

This work was supported by **National Key R & D Program of China (2023YFC3403200)** and **Shenzhen Science and Technology Program (20220814183301001)**.

References

[1] J. Luo, X. Wang, Y. Zou, L. Chen, W. Liu, W. Zhang, and S. C. Li, "Quantitative annotations of t-cell repertoire specificity," *Briefings in Bioinformatics*, vol. 24, no. 3, p. bbad175, 2023.
[2] G. Stelzer, N. Rosen, I. Plaschkes *et al.*, "The genecards suite: from gene data mining to disease genome sequence analyses," *Current Protocols in Bioinformatics*, vol. 54, no. 1, pp. 1.30.1–1.30.33, 2016.
[3] G. Song, T. Feng, R. Zhao, Q. Lu, Y. Diao, Q. Guo, Z. Wang, Y. Zhang, L. Ge, J. Pan *et al.*, "Cd109 regulates the inflammatory response and is required for the pathogenesis of rheumatoid arthritis," *Annals of the Rheumatic Diseases*, vol. 78, no. 12, pp. 1632–1641, 2019.
[4] S. Censi, C. Mian, and C. Betterle, "Insulin autoimmune syndrome: from diagnosis to clinical management," *Annals of Translational Medicine*, vol. 6, no. 17, 2018.
[5] M. Lin, Y. Chen, J. Ning *et al.*, "Insulin autoimmune syndrome: A systematic review," *International Journal of Endocrinology*, 2023.
[6] T. Ogawa, Y. Hirohashi, A. Murai, T. Nishidate, K. Okita, L. Wang, Y. Ikehara, T. Satoyoshi, A. Usui, T. Kubo *et al.*, "St6galnac1 plays important roles in enhancing cancer stem phenotypes of colorectal cancer via the akt pathway," *Oncotarget*, vol. 8, no. 68, p. 112550, 2017.

*Corresponding author: Shuai Cheng Li (shuaicli@cityu.edu.hk) and Xikang Feng (fxk@nwpu.edu.cn)
+These authors contributed equally to this work

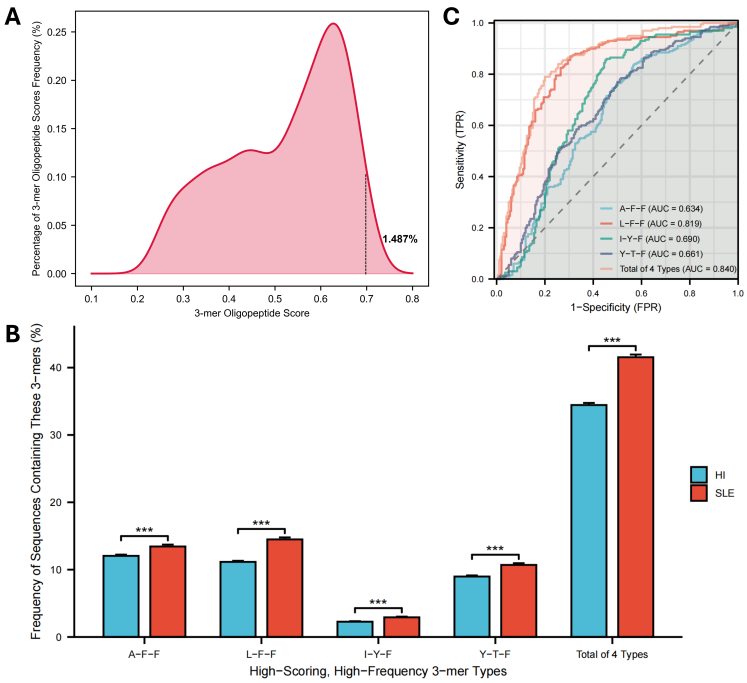

Fig. 1. **Screening of essential 3-mer oligopeptides and their role in enhancing diagnostic discrimination for TCRβ CDR3.** (A) Smoothed histogram reflects the frequency distribution of 3-mer oligopeptide scores in SLE patients, where a mere fraction, less than 1.5%, achieves a high score of 0.7 or above. (B) Clustered bar chart showing the significant frequency differences of essential 3-mer oligopeptides (3-mer) in TCR samples from SLE patients compared to HI ($p < 0.001$). (C) ROC curve showing good performance in diagnosing and classifying HI and SLE patients based on the frequency of essential 3-mer oligopeptides.

TABLE I
POTENTIAL ANTIGENS AND GENES FOR SLE UNEARTHED FROM SIGNIFICANT SEQUENCES IDENTIFIED BY DEEP LEARNING

Potential Antigen	**Potential Gene**	**Autoimmune-Related Disease**
CD109 antigen	CD109	Rheumatoid Arthritis
Insulin	INS	Type 1 Diabetes
Alpha-N-acetylgalactosaminide alpha-2,6-sialyltransferase 3	ST6GALNAC1	Colitis
Protein NPAT	NPAT	Ataxia-Telangiectasia
Islet-specific glucose-6-phosphatase-related protein	IGRP	Type 1 diabetes

[7] M. Hugonnet, P. Singh, Q. Haas, and S. von Gunten, "The distinct roles of sialyltransferases in cancer biology and onco-immunology," *Frontiers in Immunology*, vol. 12, p. 799861, 2021.

[8] D. S. Pisetsky, "Pathogenesis of autoimmune disease," *Nature Reviews Nephrology*, vol. 19, no. 8, pp. 509–524, 2023.

[9] J. Yang, N. A. Danke, D. Berger, S. Reichstetter, H. Reijonen, C. Greenbaum, C. Pihoker, E. A. James, and W. W. Kwok, "Islet-specific glucose-6-phosphatase catalytic subunit-related protein-reactive cd4+ t cells in human subjects," *The Journal of Immunology*, vol. 176, no. 5, pp. 2781–2789, 2006.

[10] S. V. Gearty, F. Dündar, P. Zumbo, G. Espinosa-Carrasco, M. Shakiba, F. J. Sanchez-Rivera, N. D. Socci, P. Trivedi, S. W. Lowe, P. Lauer *et al.*, "An autoimmune stem-like cd8 t cell population drives type 1 diabetes," *Nature*, vol. 602, no. 7895, pp. 156–161, 2022.