

Supplementary Material

Validation of miRNA signatures for ovarian cancer earlier detection in the pre-diagnosis setting using machine learning approaches

Konrad Stawiski^{1*}, Renée T. Fortner^{2,3*}, Luca Pestarino^{2,4*}, Sinan U Umu², Rudolf Kaaks³, Trine B. Rounge^{2,5}, Kevin M. Elias^{6,7}, Wojciech Fendler^{1,8**}, Hilde Langseth^{2,9**}

*contributed equally and should be considered co-first authors **contributed equally and should be considered co-last authors

¹Department of Biostatistics and Translational Medicine, Medical University of Lodz, Lodz, Poland

²Department of Research, Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway

³Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany

⁴Department of Gynecological Oncology, Division of Cancer Medicine, Oslo University Hospital, Oslo, Norway

⁵Centre for Bioinformatics, Department of Pharmacy, University of Oslo, Oslo, Norway

⁶Division of Gynecologic Oncology, Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Boston, MA, United States

⁷Harvard Medical School, Boston, MA, United States. ⁸Department of Radiation Oncology, Dana-Farber Cancer Institute, Boston, MA, United States

⁹Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK

* Correspondence:

Konrad Stawiski, M.D., Ph.D. Department of Biostatistics and Translational Medicine Medical University of Lodz Mazowiecka 15, 92-215 Lodz, Poland +48 42 42 272 53 85 konrad.stawiski@umed.lodz.pl

Renée Turzanski Fortner, Ph.D. Cancer Registry of Norway Norwegian Institute of Public Health Postbox 5313 Majorstuen 0304 Oslo, Norway +47 22 92 88 02 rtfo@kreftregisteret.no

Supplementary Figures and Tables

Supplementary Figures



Supplemental Figure 1. The principal component analysis showed moderate differences due to batch effect in RNA profiles between ovarian cancer cases and controls in the JanusRNA study.



Supplemental Figure 2. The initial principal component analysis showed significant differences in RNA profiles by study when comparing the data from Elias 2017 and the JanusRNA study.



Supplemental Figure 3. Evaluation of miRNA set from dataset described in Keller et al. (2017) in JanusRNA ovarian cancer case set. (a) miRNA counts for the evaluated panel, by time between blood collection and diagnosis (<2 years, \geq 2 to 7 years; microRNAs log10 transformed for visualization and zero values omitted from the plot); (b) Boxplots of model performance for the primary evaluated miRNA signature and random miRNA set; (c) Feature importance for each of the included miRNAs (d) Misclassification plot of the test samples based on the miRNA signature, y axis represents prediction probability for a given observation to have been sampled within 2 years prior to diagnosis.



Supplemental Figure 4. ROC curves for the XGBoost model on the evaluated miRNA sets. Each curve represents one of the five training/test split performed. The average AUC and relative 95% confidence intervals are also shown for each of the miRNA sets. In the random miRNA sets figure, each curve represents the average of each of the five training/test split repetitions curve and its \pm one standard deviation from the mean.



miRNA random sets AUC comparison

miRNA random sets feature importance В. Random set 1 Random set 2 Random set 3 miR-5189-5p -miR-9985 -miR-642b-5p -miR-6788-5p -miR-5682 -miR-548f-5p -miR-5088-5p -miR-5088-5p -miR-451b -miR-451b -miR-4494 -miR-1827 miR-6745 -miR-7160-3p -miR-6888-3p -miR-6865-5p -miR-6722-3p -miR-6722-3p -miR-6088 -miR-6088 -miR-5572 -miR-500-5p -miR-5010-3p miR-98-3p miR-6872-3p miR-6872-3p miR-6771-5p miR-687 miR-5589-3p miR-4714-5p miR-4650-5p miR-4515 miR-4318 miR-5072 -miR-5000-5p -miR-4695-3p -miR-4640-3p -miR-3689d -miR-10394-5p miR-1827 miR-4318 let-7c-3p -0.00 0.25 0.50 0.75 1.00 0.0 0.1 0.2 0.3 0.4 0.5 0.00 0.25 0.50 0.75 Random set 4 Random set 5 miR-516b-5p -miR-6780a-5p -miR-548am-3p -miR-6870-5p -miR-6133 let-7d-3p -miR-939-5p -miR-380-5p -miR-3191-3p -miR-6875-5p miR-6133 - miR-6133 - miR-4741 - miR-891a-5p - miR-6825-3p - miR-4644 - miR-3677-3p - miR-32b-3p - miR-137-3p - miR-1204 miR-0875-50 miR-1291 -miR-937-3p -miR-7108-5p -miR-6796-3p -miR-6730-5p miR-0730-3p -miR-5580-5p -miR-4774-5p -miR-4642 -miR-296-5p -0.0 0.1 0.2 0.3 0.4 0.0 0.2 0.4 0.6

Supplemental Figure 5. Boxplots of model performance for the five sets randomly selected miRNAs (14 random miRNAs for each set).

Α.



Supplemental Figure 6. Spearman correlation heatmap between the miRNA set from Elias et al. (2017) and the randomly selected miRNA with non-zero counts modelled in Supplemental Figure 4.



Supplemental Figure 7. Standardized log₁₀-transformed RPM miRNA counts (z-score of log₁₀ transformed values) dynamics from dataset described in Elias et al. (2017) in JanusRNA ovarian cancer case set.