Urinary Biomarkers for Disease Detection

Alexandre Matov[1, †]

[1] DataSet Analysis LLC, 155 Jackson St, San Francisco, CA 94111

[†] Corresponding author:

email: matov@datasetanalysis.com

Key words: Small RNA, Longitudinal Analysis, Disease Detection, Early Disease

ABSTRACT

The current healthcare system relies largely on a passive approach toward disease detection, which typically involves patients presenting a "chief complaint" linked to a particular set of symptoms for diagnosis. Since all degenerative diseases occur slowly and initiate as changes in the regulation of individual cells within our organs and tissues, it is inevitable that with the current approach to medical care we are bound to discover some illnesses at a point in time when the damage is irreversible and meaningful treatments are no longer available.

There exist organ-specific sets (or panels) of nucleic acids, such as microRNAs (miRNAs), which regulate and help to ensure the proper function of each of our organs and tissues. Thus, dynamic readout of their relative abundance can serve as a means to facilitate real-time health monitoring. With the advent and mass utilization of next-generation sequencing (NGS), such a proactive approach is currently feasible. Because of the computational complexity of customized analyses of "big data", dedicated efforts to extract reliable information from longitudinal datasets is key to successful early detection of disease.

Here, we present our preliminary results for the analysis of healthy donor samples and drug-naïve lung cancer patients.

INTRODUCTION

The cancer incidences and mortality rates worldwide demonstrate that for some cancers, such as lung, stomach, liver, esophagus, leukemia and pancreas cancer, the diagnosis is almost inevitably linked to a loss of life. The 5-years survival rate for patients with lung cancer, the "biggest killer", is 47% for stage I and can be as low as 2% for stage IV, which underscores the importance of early detection (World Health Organization, 2017). Methods for the early detection of disease based on the analysis of body fluids have become popular, however, the identification of biomarkers in urine is one of the few non-invasive

approaches. At least 1,917 human miRNAs have been described in the literature (Kozomara and Griffiths-Jones, 2014), and we have detected 947 of them in urine. The levels of miRNA change as a function of age and systematic differences exist between the two sexes (Ben-Dov et al., 2016). There are three main categories of degenerative diseases - cardiovascular (e.g., hypertension, coronary disease, myocardial infarction), neoplastic (tumors and cancers), and nervous system-related (e.g., Alzheimer's disease and Parkinson's disease). The etiology of these diseases typically involves some combination of aging and poor lifestyle choices. It is, therefore, conceivable that the collection and analysis of urinary miRNA panels longitudinally would allow to delineate changes in physiology associated with a particular disease in its early (asymptomatic) stage.

COMPUTATIONAL ANALYSIS

Figure 1 shows an example of the miRNAs detection in the urine of a healthy individual. Not all of the miRNAs we detected in urine are present in all subjects for each of the time points. We aimed at

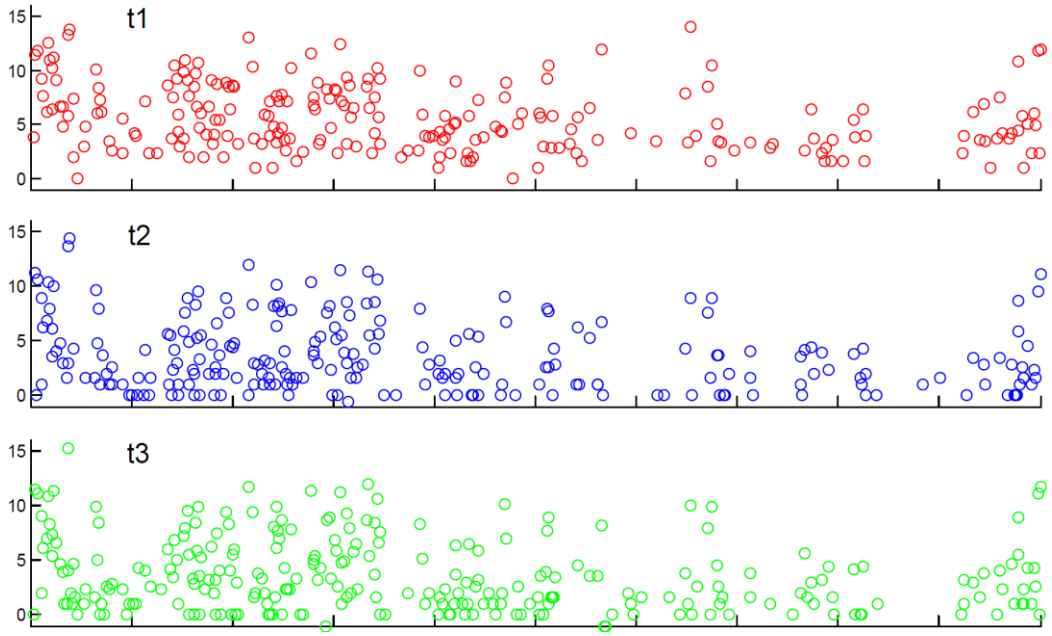

Figure 1. Expression levels (logarithmic scale) of 947 miRNAs sequenced by NGS from urine samples for one of the 15 healthy individuals over three time points. Example of visualization of urine data based on 947 miRNAs plotted on the X-axis versus miRNA expression levels (in arbitrary units) as plotted on the Y-axis. Time point 1 levels are displayed in red, time point 2 levels are displayed in blue and time point 3 levels are displayed in green.

discovering personalized biomarker panels, which would allow anyone to make informed decisions and alert about the need for lifestyle changes or to seek medical advice. In this context, we sought to identify meaningful patterns and predictive trends in miRNA levels present in patient urine and blood specimens. First, we performed a literature search and identified a panel of 25 high fidelity lung cancer biomarkers, based on miRNAs each previously linked to lung cancer development, progression and drug resistance in multiple (between three and eight) papers on the analyses of blood of lung cancer patients and primary lung tumors. Lung cancer is the leading cause of cancer-related deaths and claims more lives each year than all other major cancers combined. Lung cancers are generally diagnosed at an advanced stage because patients lack symptoms in the early stages of the disease. We measured an increase in the levels of 16 of these 25 miRNAs for one of the healthy individuals in our baseline cohort (Fig. 2). The most commonly

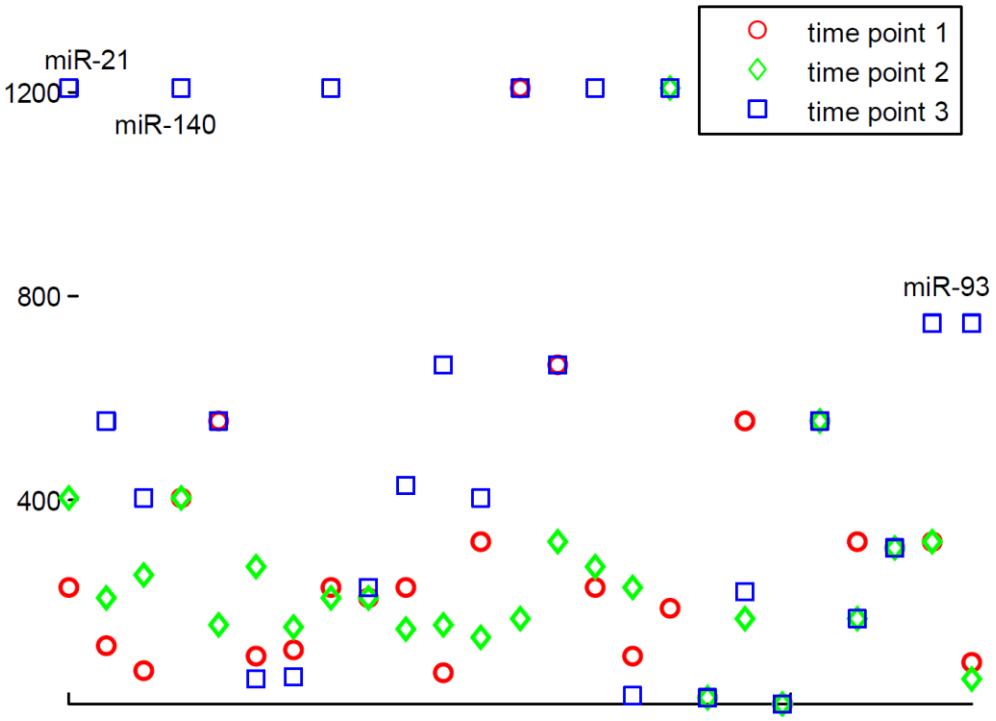

Figure 2. Longitudinal changes of lung cancer-related urinary miRs in a healthy individual. An example of up-regulation of 16 urinary biomarkers within a panel of 25 high fidelity lung cancer biomarkers, based on miRNAs each previously linked to lung cancer development, progression and drug resistance in multiple (between three and eight) papers on the analyses of blood of lung cancer patients and primary lung tumors. X-axis, list of the biomarkers: from left to right: (1) miR-21-3p, (2) miR-21-5p, (3) miR-140-3p, (4) miR-140-5p, (5) miR-155, (6) miR-200b-3p, (7) miR-200b-5p, (8) miR-223-3p, (9) miR-223-5p, (10) miR-221-3p, (11) miR-221-5p, (12) miR-145-3p, (13) miR-145-5p, (14) miR-150-3p, (15) miR-150-5p, (16) miR-200a-3p, (17) miR-200a-5p, (18) miR-205-3p, (19) miR-205-5p, (20) miR-210-3p, (21) miR-210-5p, (22) miR-339-3p, (23) miR-339-5p, (24) miR-93-3p, (25) miR-93-5p. Y-axis, abundance levels; we performed data normalization using the quantile normalization method recommended for next-generation sequencing data based on single color experiments.

published biomarkers of lung cancer, which we found to have an increase longitudinally are miRNA-21-3p, miRNA-140-3p and miRNA-93-3p. We did not consider those miRNAs for which there have been reports in the literature of a downregulation in disease, because their levels might appear decreased in urine for other physiological reasons than cancers. After establishing that it was feasible to identify published cancer miRNAs in urine, we sought to investigate whether healthy donors retain similar patterns in their miRNA profiles longitudinally via principal component analysis (Fig. 3). To this end, we processed samples of 15 healthy individuals collected every two months for three time points, i.e., 45 samples in total. This analysis showed that even if there is a certain level of variability, (i) each three

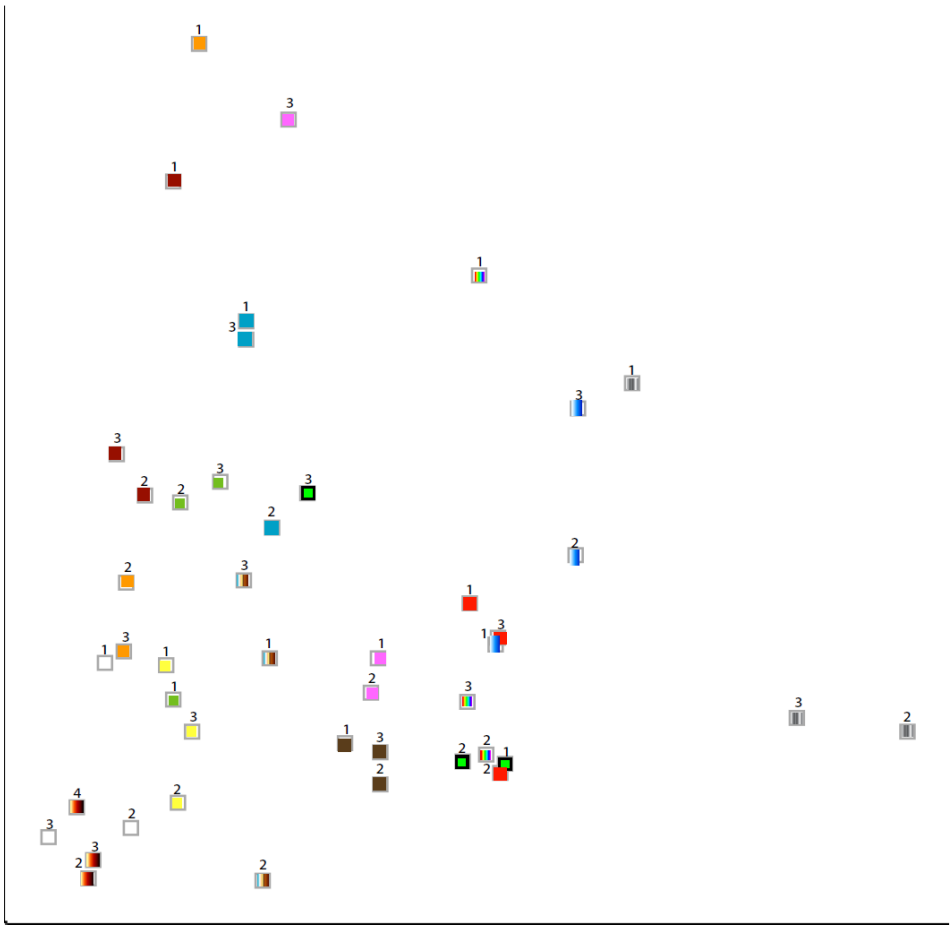

Figure 3. Longitudinal analysis of 45 samples from 15 healthy donors; X-axis Component 1 (21.4%), Y-axis Component 2 (13.1%). PCA analysis of longitudinal data of healthy donors taken at three different time points demonstrates the transient nature of the miRNAs levels. The panel consists of 15 healthy donors which provided void urine samples in the Netherlands. Each triplet 1-2-3 in the same color on the PCA scatter plot belong to the same healthy individual over three time points. The time step at which these samples were collected and sequenced was two months. One of the triplets is labeled 2-3-4 because the first sample was not processed correctly and that required the collection of another one at a later, fourth, time point.

longitudinal samples from the same healthy individual cluster together - at least for two of the three time points - and (ii) the longitudinal sample triplets for each of the different healthy individuals form separate clusters (see on Fig. 3 sample triplets per individual color-coded in the same color – pink for donor #1, red for donor #2, cyan for donor #3, blue for donor #4, brown for donor #5, yellow for donor #6, orange for donor #7, etc.). This demonstrates that, in normal physiology, urinary miRNA panels can identify the same person longitudinally.

One of our aims has been to discover personalized disease-specific miRNAs biomarker panels based on significant changes in organ or tissue regulation in disease. As a first step in this regard, we aimed at discovering population-based panels of biomarkers. The small sizes of our preliminary datasets preclude utilization of standard methods, e.g., differential expression analysis (Robinson et al., 2010), so we opted to perform information theory-based computation to identify biomarkers. As the size of our datasets increases, so will our ability to make meaningful biomarker identifications. For the selection of disease-specific panels, several approaches are available for feature selection, for instance, by maximization of mutual information (Jiao et al., 2015) or applying the "maximum of the minimum criterion" (Bennasar M., 2015). The preliminary methodology we utilized for computation of relative entropy via Kullback-Leibler (KL) divergence is based on an adaptive minimax rate-optimal estimator (Han et al., 2016) of the changes in disease from healthy state(s) to cancerous lesion(s) and malignant tumor(s). Consider the KL divergence as:

$$D(P||Q) \triangleq \begin{cases} \sum_{i=1}^{S} p_i ln \frac{p_i}{q_i} \; if \; P \ll Q, \\ +\infty \qquad \text{otherwise}, \end{cases} \qquad (1)$$

where two patient cohorts are considered, $P = \{p_1, \dots, p_s\}$ and $Q = \{q_1, \dots, q_s\}$, over a common set of miRNAs of length S (S=947 miRNAs for this dataset). Testing this approach for significant feature selections on a small cohort of 13 stage IV lung cancer patients' urine samples, we selected miRNAs by computing Eq. (1). We set our threshold at 1.2 bits divergence empirically and this allowed us to identify

20 biomarkers discriminative of lung cancer (Fig. 4). All 20 biomarkers have previously been published in the literature on lung cancer based on analyses of primary lung tumors and blood from lung cancer patients (Bao et al., 2018; Wan and Zheng, 2021; Wang et al., 2014). This selection demonstrates the suitability and ability of this method for the identification and selection of disease-specific biomarkers. Even if this computation is based on small sample cohorts and it is not patient-specific, it indicates the possibility to detect lung cancer in urine samples. Our objective will be to provide every individual with the option to make data-driven decisions in the context of the prevention of the progression of degenerative diseases.

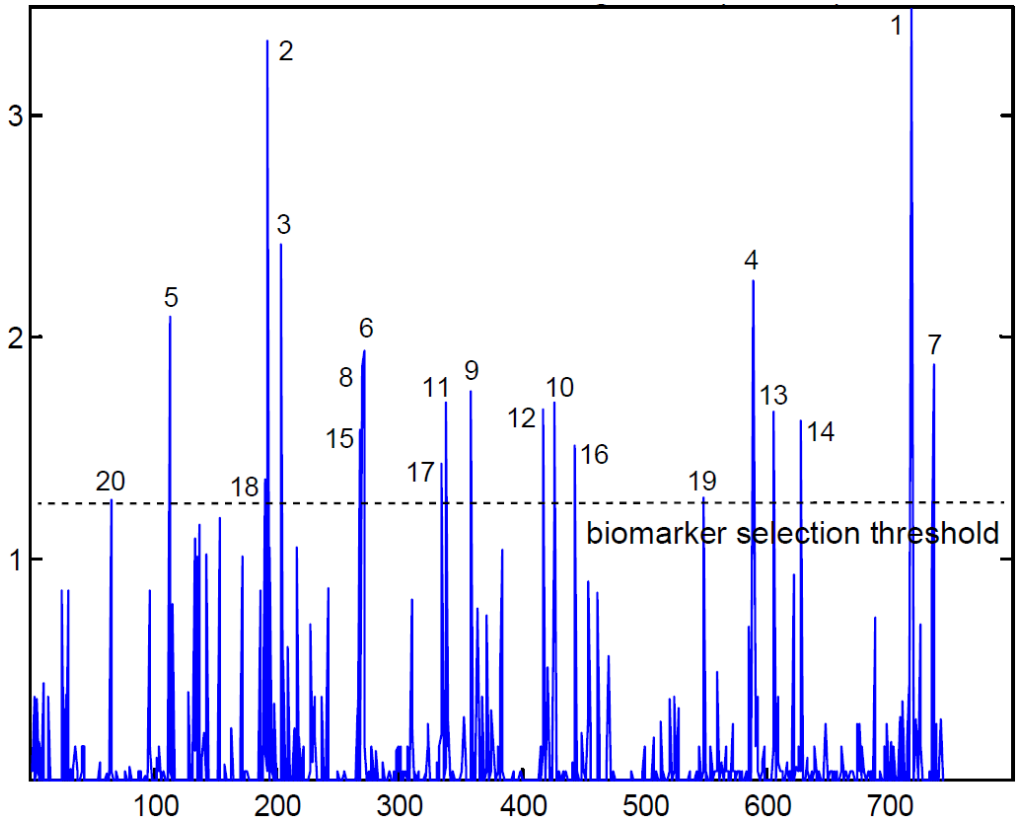

Figure 4. Kullback-Leibler divergence between drug-naïve lung cancer patients and healthy individuals. miRs levels in a cohort of 28 participants (13 stage IV drug-naïve lung cancer patients and 15 healthy individuals) allowed for the selection of a panel of 20 discriminative biomarkers out of 947 detected by next-generation sequencing in urine samples: (1) miR-891a-5p, (2) miR-196a-5p, (3) miR-200a-5p, (4) miR-577, (5) miR-141-3p, (6) miR-29c-3p, (7) miR-95-3p, (8) miR-29b-3p, (9) miR-361-5p, (10) miR-429, (11) miR-335-5p, (12) miR-421, (13) miR-628-3p, (14) miR-660-5p, (15) miR-29a-3p, (16) miR-4454, (17) miR-330-3p, (18) miR194-5p, (19) miR-532-3p, (20) miR-1271-5p. All 20 biomarkers have previously been published in the literature on lung cancer based on analyses of primary lung tumors and blood from lung cancer patients.

CONCLUSIONS

The cancer incidences and mortality rates worldwide demonstrate that for some cancers, such as lung, stomach, liver, esophagus, leukemia and pancreas cancer, the early diagnosis is of critical importance. We will develop algorithmic solutions to facilitate the longitudinal analyses of transcriptomic miRNA datafrom tens of thousands of patients (about 1,917 miRNAs have been described in the literature and we have detected 947 of them in urine) and the early detection of disease. To achieve this end, we will use multivariate data and multi-dimensional clustering analyses as well as information theory approaches based, for instance, on entropy computation and feature selection using joint mutual information maximization. Our overall objective is to significantly improve the quality of life in the last three to four decades of living.

The methodology we propose may allow the development of molecular diagnostics tests based on nucleic acid markers for monitoring of cardiovascular, neoplastic, and diseases of the nervous system based on the longitudinal analyses of body fluids. Further, the approach may provide novel ways for treatment evaluation. Correlation between changes in biomarker levels and treatment responses will allow for the early detection of a lack of response after treatment and ultimately for optimal drug selection. It will also facilitate the discovery of biomarkers for the prediction of disease relapse. In the long run, longitudinal analysis of nucleic acids will allow for the development of novel targeted drugs. Recent literature has described many examples of miRNAs that could be targeted in disease (Duygu et al., 2016; Kim et al., 2017; Singh and Sen, 2017). Therefore, monitoring of their levels in healthy individuals and patients undergoing disease treatment will likely provide valuable datasets for the pharmaceutical industry as well as for practicing physicians, ultimately allowing them to select the most efficacious treatment sequence and drug combination for each patient.

MATERIALS AND METHODS

Sample Processing

Small RNA were extracted from de-identified urine samples by Norgen Biotek Corp., Thorold.

Data Analysis

All data analysis programs and graphical representation of the results were developed in R and Matlab. The computer code is available for download at: https://github.com/amatov/DiseaseDetectionUrine.

Ethics Declaration

IRB of the Institute of Regenerative and Cellular Medicine.

Ethical approval was given.

ACKNOWLEDGEMENTS

I thank JR&D Services, Rotterdam for funding in the context of NGS data. I am grateful to the Institute of Regenerative and Cellular Medicine, Santa Monica for issuing the Institutional Review Board protocol approval IRCM-2019-201, IRB DS-NA-001 for the observational study “Longitudinal analysis of next-generation sequencing of nucleic acids for early detection of degenerative diseases such as cardiovascular, neoplastic and diseases related to the nervous system” and James Faber for his feedback regarding the protocol and the process of approval.

REFERENCES

World Health Organization. 2017. Global yearly cancer incidence and mortality. *American Cancer Society*.

Bao, M., S. Pan, W. Yang, S. Chen, Y. Shan, and H. Shi. 2018. Serum miR-10a-5p and miR-196a-5p as non-invasive biomarkers in non-small cell lung cancer. *International journal of clinical and experimental pathology*. 11:773-780.

Ben-Dov, I.Z., V.M. Whalen, B. Goilav, K.E. Max, and T. Tuschl. 2016. Cell and Microvesicle Urine microRNA Deep Sequencing Profiles from Healthy Individuals: Observations with Potential Impact on Biomarker Studies. *PLoS One*. 11:e0147249.

Bennasar M., H.Y., Setchi R. 2015. Feature Selection Using Joint Mutual Information Maximization. *Expert Systems With Applications*. 42:8520-8532.

Duygu, B., L.J. de Windt, and P.A. da Costa Martins. 2016. Targeting microRNAs in heart failure. *Trends Cardiovasc Med*. 26:99-110.

Han, Y., J. Jiao, and T. Weissman. 2016. Minimax Rate-optimal Estimation of KL Divergence between Discrete Distributions. *Int Symp Inf Theory Appl*. 2016:256-260.

Jiao, J., K. Venkat, Y. Han, and T. Weissman. 2015. Minimax Estimation of Functionals of Discrete Distributions. *IEEE Trans Inf Theory*. 61:2835-2885.

Kim, D., H.R. Chang, and D. Baek. 2017. Rules for functional microRNA targeting. *BMB Rep*. 50:554-559.

Kozomara, A., and S. Griffiths-Jones. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. 42:D68-73.

Robinson, M.D., D.J. McCarthy, and G.K. Smyth. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*. 26:139-140.

Singh, A., and D. Sen. 2017. MicroRNAs in Parkinson's disease. *Exp Brain Res*. 235:2359-2374.

Wan, N., and J. Zheng. 2021. MicroRNA-891a-5p is a novel biomarker for non-small cell lung cancer and targets HOXA5 to regulate tumor cell biological function. *Oncology letters*. 22:507.

Wang, D., J. Gu, T. Wang, and Z. Ding. 2014. OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics (Oxford, England)*. 30:2237-2238.