BIOGNOSYS
NEXT GENERATION PROTEOMICS

Biognosys AG
Wagistrasse 21
8952 Schlieren (Zurich)
Switzerland

t: +41 (0) 44 738 20 40
f: +41 (0) 44 738 20 49
info@biognosys.com
www.biognosyscom

Big Proteomic Data from a Digital Biobank Boost the Confidence in Biomarker Discovery

Andrés Lanzós, Roland Bruderer, Jan Muntel, Lukas Reiter, Oliver Rinner and Sebastian Schegk

Introduction

Despite massive research efforts during the last decades, FDA's Biomarker Qualification Program has only qualified 16 biomarkers so far [1]. In addition, less than 60% of biomarker studies end up published within 18 months of completion [2]. The main reasons why biomarker discovery is failing to the promise of the omics era include: overestimated effect sizes due to low sample size, presence of false positives due to lack of validation studies, and missed associations of candidate biomarkers to confounding factors [3]. Here we present a pilot digital biobank as a solution to mitigate such limitations. By adding big proteomic data to their analysis, researchers can instantly validate candidate biomarkers in bigger sample sizes and search for associations to other conditions.

Methods

For this pilot, we gathered the plasma proteome measurements of previously acquired 476 subjects from the DiOGenes study [4], where samples were taken before and after a weight loss program. In short, we applied the following workflow for sample measuring and data generation:

1. LC-MS/MS measurement with a Waters M-Class LC instrument coupled to a Thermo Scientific Orbitrap Fusion Lumos MS instrument.
2. DIA identification and quantification of more than 550 proteins with Spectronaut Pulsar X.
3. Data normalization and batch effect correction with a tailormade R pipeline using global median normalization and batch mean centering.
4. Storing of protein abundances and clinical metadata in an SQLite database.

Preliminary Data

Having generated a pilot digital biobank, we then designed several applications of big proteomics data to boost biomarker discovery. Here we show one of them, where a researcher can assess how many candidate biomarkers of weight loss are likely true and false positives when validated with the complete biobank.

Biognosys AG
Wagistrasse 21
8952 Schlieren (Zurich)
Switzerland

t: +41 (0) 44 738 20 40
f: +41 (0) 44 738 20 49
info@biognosys.com
www.biognosyscom

First, we collected all 450 and 464 samples before and after weight loss, respectively. Then, we used the Limma method to test for differential abundance of proteins between the two conditions. Finally, we defined the top 10 proteins sorted by p-value as “biobank candidates”.

We repeated the same process for 200 simulated researchers with 50 samples per condition randomly selected from all samples, defining the top 10 proteins by p-value as “researcher candidates” in each simulation. Then, we measured the percentage of “researcher candidates” in each simulation that are also “biobank candidates”, i.e., likely true positives.

This analysis indicates that with 50 samples per condition, a researcher would have on average 70% likely false positives among his/her top 10 candidates. Thus, by focusing on the 30% of likely true positives, the researcher can significantly narrow down further experiments and increase the success rate of his/her biomarker discovery.

This is just one of the many applications of a digital biobank to boost biomarker discovery through big proteomic data analyses. A digital biobank with more samples and clinical conditions would allow more possibilities, including:

1. Increased statistical power: combination of researcher’s samples with those from the biobank for group comparisons like tumor against healthy.

2. Biomarker condition specificity: detection of biomarkers of interest in other conditions not initially tested, such as tumor biomarkers common between adenocarcinoma (researcher) and squamous cell carcinoma (biobank).

3. In silico only studies: biomarker discovery using only samples from the digital biobank without measuring new ones.

Novelty aspect

Analysis of big proteomic data from a digital biobank can reduce on average 70% likely false positives in biomarker discovery.

References

[1] U.S. Food & Drug Administration, “List of Qualified Biomarkers.” https://www.fda.gov/drugs/biomarker-qualification-program/list-qualified-biomarkers (accessed Jun. 23, 2021).

[2] J. P. A. Ioannidis and P. M. M. Bossuyt, “Waste, leaks, and failures in the biomarker pipeline,” *Clin. Chem.*, vol. 63, no. 5, pp. 963–972, 2017, doi: 10.1373/clinchem.2016.254649.

[3] J. P. A. Ioannidis and O. A. Panagiotou, “Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses,” *JAMA - J. Am. Med. Assoc.*, vol. 305, no. 21, pp. 2200–2210, 2011, doi:

Biognosys AG
Wagistrasse 21
8952 Schlieren (Zurich)
Switzerland

t: +41 (0) 44 738 20 40
f: +41 (0) 44 738 20 49
info@biognosys.com
www.biognosyscom

10.1001/jama.2011.713.

[4] R. Bruderer *et al.*, “Analysis of 1508 plasma samples by capillary-flow data-independent acquisition profiles proteomics of weight loss and maintenance,” *Mol. Cell. Proteomics*, vol. 18, no. 6, pp. 1242–1254, 2019, doi: 10.1074/mcp.RA118.001288.