

Big Data and Host-Pathogen Interactions

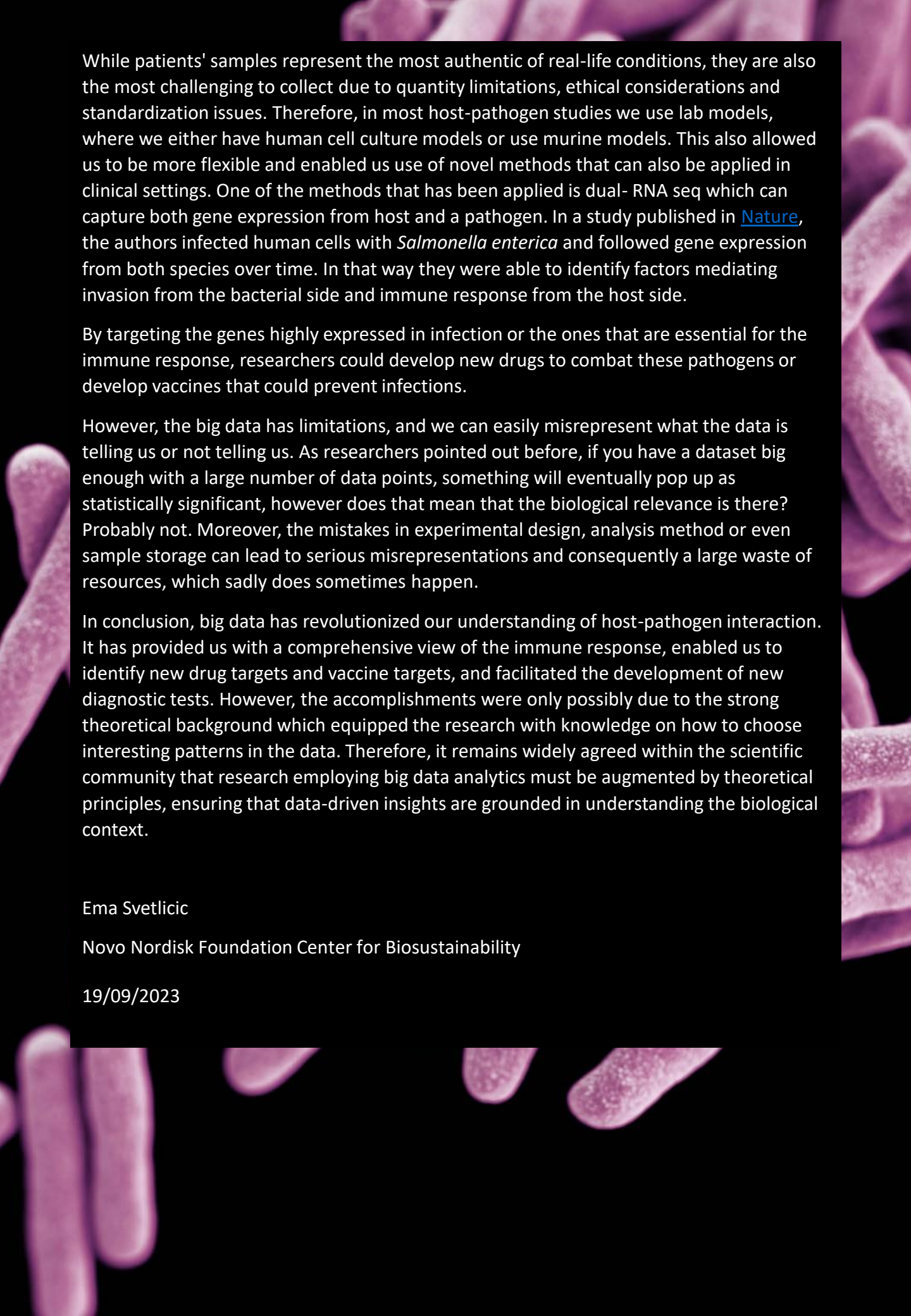
As the world battles against emerging and re-emerging infectious diseases, understanding the interactions between host and pathogen becomes increasingly critical. The human immune system is complex, and pathogens like bacteria and viruses have developed various strategies to evade it. But with the advent of big data, we have a powerful tool to decode the host-pathogen interplay, paving the way for new diagnostic, therapeutic, and preventive strategies.

When we talk about big data it usually means that research is data-driven which is essentially applying analytical techniques and computational power to explore and extract interesting features in the data. Contrary to the data driven research we have a theory-driven research- a more traditional approach in which we use the existing knowledge and careful observations to develop a hypothesis. Then the hypothesis is tested and validated usually by investigating a particular pathway or molecule. However, when we look at the complexity of the host-pathogen interactions, sometimes there is not a strong theoretical base for us to develop a hypothesis. This is when the big data comes in light. So instead of being focused on a few interactors and a single pathway we are trying to reveal a comprehensive view. It can help us understand how pathogens are interacting with the host immune defense system, or what previously unknown pathways are activated in the pathogen when trying to survive in the host environment.

But how is that big data generated? Well, usually through so called omics-based technologies. The meaning of word omics is “a study of the totality of something”. So this is precisely what we are trying to do in biological setting. For example we are trying to study the structure and function of the total genetic content in an organism and then we are calling it genomics. If we are trying to do the same with proteins then we call it proteomics, metabolites- metabolomics and so on. You get the gist.

How does that look in practice? In a recent study published in [Cell Systems](#), scientists sampled human plasma from patients diagnosed with COVID-19 and followed how the plasma protein abundances are changing over time and with the disease progression. With the generated proteome data they were able to identify prognostic biomarkers of the disease, meaning if the target protein is elevated in the early phase, it could be a sign of future clinical deterioration.

[Another study](#), in this case concerning bacterial infection, published in the journal *Cell*, researchers analyzed the gene expression profiles and metabolites of patients infected with the bacterium *Staphylococcus aureus*. They identified a set of genes whose expression levels could predict whether the patient would develop a severe infection or not. The latter is an example of how several omics technologies are used to tackle a specific problem in an approach, we call multi-omics research.

The background of the slide is a microscopic image of purple rod-shaped bacteria, likely Salmonella enterica, arranged in various orientations and some showing flagella. The bacteria are set against a dark background, making their purple color stand out.

While patients' samples represent the most authentic of real-life conditions, they are also the most challenging to collect due to quantity limitations, ethical considerations and standardization issues. Therefore, in most host-pathogen studies we use lab models, where we either have human cell culture models or use murine models. This also allowed us to be more flexible and enabled us use of novel methods that can also be applied in clinical settings. One of the methods that has been applied is dual- RNA seq which can capture both gene expression from host and a pathogen. In a study published in [Nature](#), the authors infected human cells with *Salmonella enterica* and followed gene expression from both species over time. In that way they were able to identify factors mediating invasion from the bacterial side and immune response from the host side.

By targeting the genes highly expressed in infection or the ones that are essential for the immune response, researchers could develop new drugs to combat these pathogens or develop vaccines that could prevent infections.

However, the big data has limitations, and we can easily misrepresent what the data is telling us or not telling us. As researchers pointed out before, if you have a dataset big enough with a large number of data points, something will eventually pop up as statistically significant, however does that mean that the biological relevance is there? Probably not. Moreover, the mistakes in experimental design, analysis method or even sample storage can lead to serious misrepresentations and consequently a large waste of resources, which sadly does sometimes happen.

In conclusion, big data has revolutionized our understanding of host-pathogen interaction. It has provided us with a comprehensive view of the immune response, enabled us to identify new drug targets and vaccine targets, and facilitated the development of new diagnostic tests. However, the accomplishments were only possibly due to the strong theoretical background which equipped the research with knowledge on how to choose interesting patterns in the data. Therefore, it remains widely agreed within the scientific community that research employing big data analytics must be augmented by theoretical principles, ensuring that data-driven insights are grounded in understanding the biological context.

Ema Svetlicic

Novo Nordisk Foundation Center for Biosustainability

19/09/2023