Health Security
Volume 22, Number 2, 2024 Mary Ann Liebert, Inc
DOI: 10.1089/hs.2023.0076

Commentary

Computational and Systems Biology Advances to Enable Bioagent Agnostic Signatures

Andy Lin,* Cameron M. Torres,* Errett C. Hobbs, Jaydeep Bardhan, Stephen B. Aley, Charles T. Spencer, Karen L. Taylor, and Tony Chiang

Keywords: Biodefense R&D, Data science, Immunology, Funding for biodefense, Bioagent agnostic signatures

Introduction

Historically, threat identification and characterization has centered on lists of specific agents known to cause severe harm to human, animal, or agricultural health, such as the Federal Select Agent Program Select Agent and Toxin list,[1] the World Health Organization Prioritization List,[2] and the National Institute of Allergy and Infectious Diseases (NIAID) Emerging Infectious Diseases and Pathogens List.[3] List-based biodefense approaches focus on agents that have been part of state-sponsored biological weapons programs or are especially dangerous known natural threats. However, previous work has highlighted the limitations of framing biodefense strategies around defined lists.[4] List-based approaches are ill-equipped to accommodate threats posed by emergent, reemergent, or novel pathogens, as demonstrated by SARS-CoV-2. While there were warnings about coronaviruses with pandemic potential,[5] SARS-CoV-2 did not appear in any lists prior to the outbreak; yet, it caused significant loss of life and disrupted social order worldwide. Early, rapid detection and characterization of unknown pathogens is an integral component of a robust biodefense posture. In addition, the ability to characterize how an unknown agent will likely affect human, animal, and plant health is a crucial requirement for biopreparedness and response.

Leiser et al[4] proposed a strategy to augment list-based approaches by characterizing threats based on how they affect human, animal, or plant hosts. Specifically, they advocated that the biodefense community should shift from an identification-based approach to a characterization-

Andy Lin, PhD, is a Linus Pauling Distinguished Postdoctoral Fellow; Errett C. Hobbs, PhD, is a Data Scientist; Karen L. Taylor, MS, is a Research Line Manager; and Tony Chiang, PhD, is a Data Scientist; all in the National Security Directorate, Pacific Northwest National Laboratory, Seattle, WA. Cameron M. Torres is a Graduate Research Assistant and Wieland Fellow, Department of Biological Sciences; Stephen B. Aley, PhD, is a Professor, Biological Sciences, and an Associate Vice President for Research, Sponsored Projects; and Charles T. Spencer, PhD, is an Associate Professor, Biological Sciences, and Edward and Barbara Brown Egbert Endowed Chair of the Department of Biological Sciences; all at the University of Texas at El Paso, El Paso, TX. Jaydeep Bardhan, PhD, is a Research Line Manager, Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA. (*co-first authors)

An earlier version of this commentary was posted as a preprint on *arXiv* on February 29, 2024 (https://doi.org/10.48550/arXiv.2310.13898).

based one by developing bioagent agnostic signatures (BASs), defined as measurable suites of biomarkers that accurately and reproducibly assess the impacts of infection or intoxication without a priori knowledge of an agent.

Retooling the US biodefense posture from a list-based approach to a dual list-based and BAS-based approach will require policy changes, technological improvements, and improved data analytics. Encouragingly, recent policy shifts have signaled how the US government recognizes the need for new technologies to counter chemical and biological threats beyond the preexisting lists of anticipated pathogens and toxins.[6] Despite this awareness, the biodefense community has only just begun to develop technologies for identifying usable BASs.

Enabling a BAS-based approach to complement existing list-based approaches to biodefense would require additional investment. Specifically, development of BASs would work in 2 dimensions: (1) pathogen characterization/classification and (2) host response characterization. While the biodefense community has started developing technologies to categorize and detect BASs, it has not addressed the data science challenges that hinder this new approach. In this commentary, we highlight promising new immunological approaches that could be leveraged for BAS development, data science problems that the community must resolve to identify BASs, and possible ways forward.

Threat Agnostic Detection Technologies

Opportunities to Infer Pathogenicity From Host Response

Currently, biological agent detection relies on screening samples against existing databases (eg, polymerase chain reaction [PCR], sequencing, or proteomics) or probing with antibodies specific for particular entities (eg, lateral flow immunoassays [LFIAs] or enzyme-linked immunosorbent assays [ELISA]).[7,8] Despite their reliability for diagnosis and epidemiology,[9-11] they unfortunately lack agent agnostic characteristics.

One promising direction for developing BASs would involve identifying signatures based on the host immune response, rather than a predetermined sequence. Detection of foreign invaders by the host's immune system relies first on the generation of binding protein receptors capable of detecting the broadest possible range of epitopes, initially nonspecific to any particular agent. After generation in the body, these proteins are subjected to negative selection that removes any self-reactive receptors. Ideally, whatever survives this selection process must recognize nonself, foreign antigens. Just as innate immune receptors identify classes of microbes through common structures, an agent agnostic detection approach might also discern structures not represented in human populations (eg, terminal α-1,3 galactosyl moieties in parasites[12]). This approach would create a new type of assay, serving as a proxy for immune cells, that would simultaneously interrogate numerous analytes and integrate data from across multiple platforms to detect BASs.

Promising Biotechnology for Bioagent Agnostic Detection

Currently, no single screening model is suitable for BAS discovery because they rely on highly specific molecules or searches against established references. Development of new technologies coupled with modifications to existing methods is needed to enable a BAS-based approach to pathogen detection. For example, creating an artificial immune system could allow for the direct testing of unknown samples (either environmental or clinical) to detect the presence of a foreign entity to which the "immune system on a chip" responds. Technologies utilizing cell responses have begun these advancements for clinical use,[13,14] yet they are still limited by the need to directly analyze selected immune factors determined from analysis of specific diseases rather than general factors indicative of a response to unknown pathogens. From the environmental sample perspective, the Defense Advanced Research Projects Agency "Friend or Foe" program has developed technologies that use cell response to differentiate between pathogenic and nonpathogenic bacteria.[15]

Recent modifications to flow cytometry, specifically the development of mass cytometry,[16,17] provide another example of how modern technologies can advance the goal of threat agnostic biodefense. While flow cytometry has been used for decades for cell analysis, its sensitivity for generating BASs is capped because of the limited number of predefined cell surface markers. An expanded number of markers at a higher specificity is now possible with mass cytometry, which also allows for the simultaneous detection of immune cells and changes in their protein expression as biomarkers for threat detection.[18,19] Mass cytometry has the potential to be adapted to monitor a wide range of analytes by creating a "multiplexed" panel of biomarkers/BASs indicative of a threat. The ability to measure multiple omics (eg, multiomics) within a single cell[20] would also serve as a powerful tool for probing multiple cell types, such as infected cells (eg, lung epithelial cells) or immune cells (eg, CD8+ cytotoxic T cells) for host response, which is important to capture pathogenic pathways that target specific cells or tissues (eg, a respiratory pathogen may have little effect on a muscle cell). These technologies provide insights that were not available in previous decades. However, as we discuss in future sections, computational analysis, such as batch corrections in cytometry and unsupervised learning for clustering heterogeneous populations of single cells, remain challenging.

Agnostic Technologies Require Collection of Multiomics Data

As implied by the examples above, the technologies that are most useful in generating BASs will simultaneously measure the same sample using multiple platforms, because analyses that result from a single technology are unlikely to fully describe the various mechanisms underlying a response (Figure 1). For instance, simply measuring inflammation can indicate a wide number of disease states, ranging from autoimmune disorders to cancer to infection. Metaanalyses of these studies show diverse ranges of values, the variance in which could indicate subdomains with the overarching designation of inflammation.[21-24] Fully understanding these inflammation measurements necessitates the integration of immunological and multiomics surveillance platforms. Defining and classifying what constitutes threat-induced immune responses will require tying omics and host response data together to distinguish the known from the unknown. We note that while BASs generated from a single omic data type can be successful, we expect that BASs generated from multiomics data are likely to be more robust and reflective of biology because multiomics data are better aligned to fully describe the various mechanisms underlying a response.

A comprehensive system for the detection of human infections would also require the simultaneous evaluation of foreign molecules and modulations in host responses. Such a system would enable BAS development from pathogen characterization, host response characterization, or both. A number of diagnostic methods in use today could be optimized to do this; for example, multianalyte LFIAs can monitor for infection in wounds by detecting interleukin 6 and pathogen DNA simultaneously, assessing both the host response and presence of a specific pathogenic agent.[9] In addition, microsphere-based flow cytometric techniques allow detection of multiple analytes within a single low-volume sample.[25] As previously noted, it is now possible to measure mRNA and protein levels (via sequencing and mass spectrometry, respectively) from a single, unique cell.[20] However, a practical challenge rests in defining which analytes represent a host-dependent threat indicator in a way that is not agent specific. More broadly, as we discuss later, analyzing data that have been jointly acquired from different modalities remains a challenge.

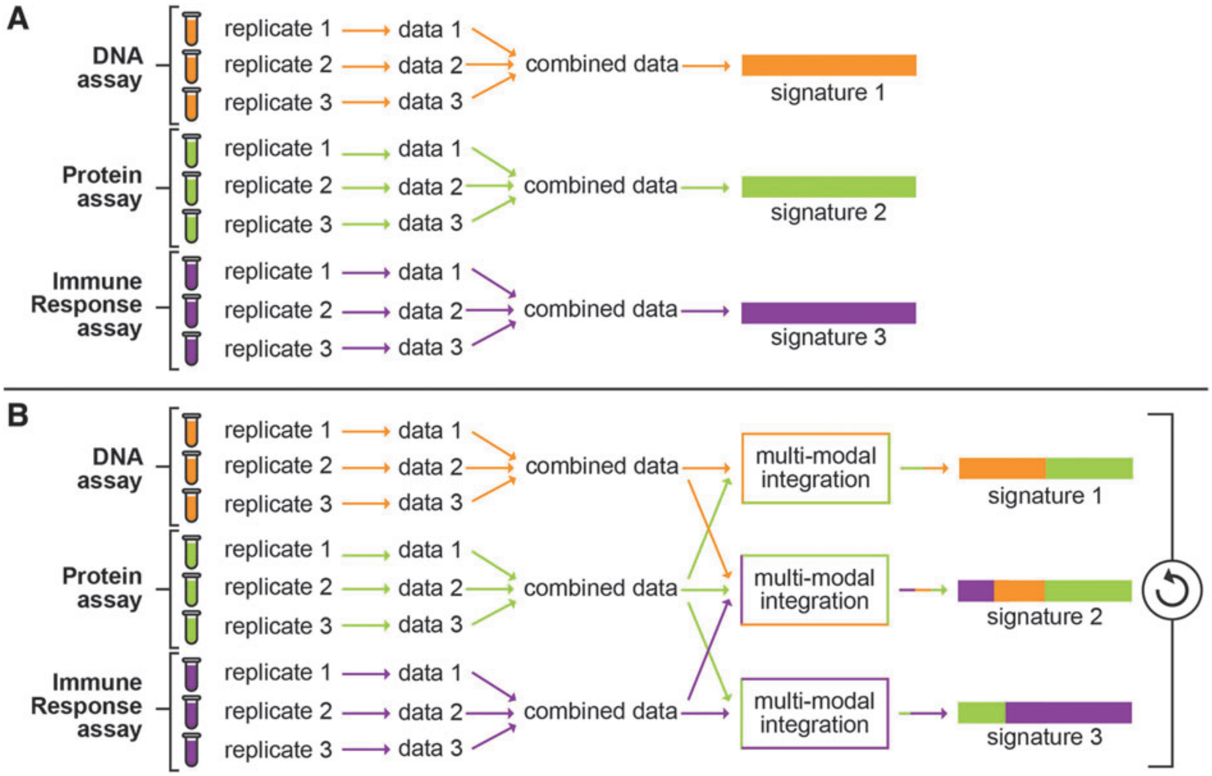

Figure 1. Comparison of traditional versus a bioagent agnostic signatures (BAS)-style approach toward signature development. (A) The traditional method for signature development. This method takes data from a single assay and generates a single signature from it. (B) Proposed approach toward development of BAS signatures. In this method, data from multiple assays are combined via multimodal integration to generate multiple signatures.

Common Gaps in Computational Biology

Absence of Baseline Information

One major gap for signature development is the lack of a consistent baseline that would allow scientists to extrapolate signatures. In order to establish capability in recognizing the presence of a threat, it is critical that we first define the composition of a healthy immune system. For example, as of 2023, no public model of a healthy generic immune system exists to facilitate comparisons against data pertaining to disease states. The nonsystematic accumulation of patient samples over the decades comes primarily from individuals with a disease phenotype and gives wide ranges of values for individual measurements. The inference of disease state necessitates access to a healthy control group. This holds true whether we are comparing genetic data among known sequenced genomes, sifting through proteomics data from environmental samples, conducting metabolomics analyses of human blood samples, or measuring cellular responses in patients. The successful development of BASs will likely require investments in healthy longitudinal cohorts to systematically probe healthy populations.

Although the biodefense community recognizes the importance of baselines and negative controls, few projects are dedicated to measuring these baselines. Programs such as All in One Breath, which aims to understand what molecules people in a healthy population exhale, represent important progress, but only a few publications have investigated similar themes[26]; additional work in this area is needed. We note that understanding baseline omics profiles is challenged by the fact that noninfectious conditions, such as aging[27] and obesity, are known to have associations with various immune markers.

Complete and Detailed Metadata Required for Signature Development

Detailed metadata annotations of the collected data are necessary for BAS development. Because data production has outgrown any individual researcher's ability to analyze data manually, scientists increasingly use algorithms to elucidate biological signals. Obtaining and identifying usable metadata are, therefore, increasingly important to ensure that analyses produce sensible outputs. Standards such as the FAIR (Findability, Accessibility, Interoperability, and Reuse of digital assets) Principles aim to normalize metadata annotation for reuse.[28] Unfortunately, while data are routinely shared, data reuse still remains a challenge due to barriers such as unstandardized metadata collection and uncoordinated data dissemination.[29] Policies that promote data sharing, such as those implemented by the National Institutes of Health in 2023,[29,30] support broader adoption of embracing FAIR Principles. If widely embraced, open sharing will unleash a wealth of data supporting the development of BASs.

Identifying BASs necessitates a metadata analytical approach, but no standardization exists that allow metadata to move beyond their association with individual datasets or analyses. Precision and consistency across studies are required in metadata analysis to prevent masking important biological differences. For example, although *prostate cancer* and *prostate cancer free* are valid metadata annotations, they are not precise because cancer subtypes may have different mechanisms and treatments. In addition, associating metadata with individual experiments, rather than with entire conclusions, supports better future use. For example, metadata on Proteomics Identification Database (PRIDE),[31,32] a public proteomics repository, is associated with entire reports rather than individual runs. This prevents future analysis of the dataset if the file naming scheme is unclear, in turn inhibiting the development of BASs.

Finally, different metadata schemes must be seamlessly interchangeable. Currently, each field has its own metadata scheme, which does not necessarily relate to the schemes of other fields. The lack of a common standard to discuss and relate datasets hinders multidisciplinary research. Additional complexity arises from the different identifiers used across different nomenclature systems. For example, the gene and protein IDs for the same entity differ across systems, such as the National Center for Biotechnology Information (NCBI), the Universal Protein Resource (UniProt), and the European Bioinformatics Institute (EBI). While these entities offer the ability to translate across nomenclature systems, using them, in practice, remains challenging. For example, the mapping across nomenclature systems can be imprecise due to many-to-many relationships (eg, a gene can be affiliated with multiple gene products). Another challenge is that these translation systems must be constantly updated as the entries in different nomenclature systems are added or updated. Creating consistent and widely used mappings between different metadata schemes would greatly enable multidisciplinary, integrative research that reflects the nature of BAS characterization and identification.

Data Harmonization

Multiomic analysis necessitates our ability to integrate disparate datasets for a unified analysis. Data integration would be simplified if experiments were designed for such an analysis a priori so as to build anchor points into each independent data subset. Historically, studies tested a single hypothesis without consideration for future use. The emergence of public data repositories in the past several decades has spurred the development of integrative methods to analyze archived data for additional biological insight.[33-36] Although there has been some success, additional investment is needed to further improve these computational techniques.

Analyzing data across multiple experiments requires careful thought due to issues associated with batch correction and normalization. Current bench science limitations require that data be generated in multiple batches, rather than all at a single time (the additional redundancy is an incidental benefit). Systematic heterogeneity both within and across batches of data (Figure 2), also known as batch effect, may confound analysis. Samples (even within batches) need to be normalized to account for heterogeneity that can arise from the sample-specific treatments essential to the study, whereas heterogeneity between batches often arises from systematic effects due to external factors of the experiment. These differences must be controlled or mitigated when analyzing data across batches. While multiple strategies exist for batch correction[37-40] and batch normalization,[41,42] these processes are labor intensive, often requiring manual oversight. Future work is needed to improve and standardize batch correction and normalization methods when combining data from the same technology (eg, flow cytometry), different technologies functionally measuring the same event (eg, microarrays and next-generation sequencing), and different technologies measuring different yet correlated events (eg, next-generation sequencing and metabolomics). Although sequence-based batch effects are discrete, continuous batch effects do occur (eg, when measuring an analyte on a mass spectrometer, its mass-to-charge ratio will drift over time).[37,38] As a result, methods that can simultaneously correct both discrete and continuous batch effects are also needed for multimodal integration.

Successful batch correction necessitates detailed metadata and sound experimental design to allow for correct stratification and prevent confounding variables from being introduced. For example, if cancer samples were sequenced at 1 facility and control samples were sequenced at a second facility, it would become impossible to disentangle the cancer signal from the processing facility unless proper control of confounders had been planned for from the beginning. Proper experimental design prior to data generation can minimize these sampling problems and provide the details needed to successfully integrate experiments into a unified meta-analysis.

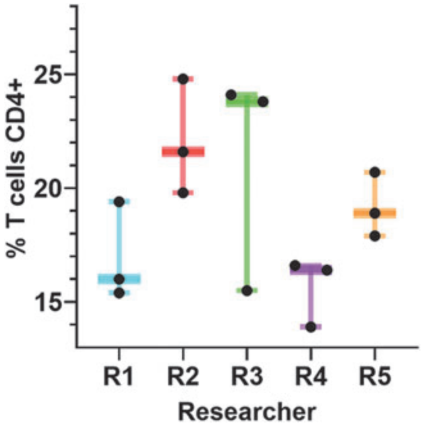

Figure 2. Experimental batch effect hinders data harmonization. Simultaneous triplicate flow cytometric analysis of the percentage of CD4+ T cells by 5 separate researchers (R1-R5) of an identical sample acquired on a Beckman Coulter Gallios flow cytometer. There is high variability of the measured percentage of CD4+ T cells within and between researchers. This variation is an example of the difficulty in determining typical versus atypical results for even well-established current assays that have been in use for decades. The data in this figure are from the laboratory of C. Spencer.

Scalable Computational Models Are Needed for Data-Centric Continual Integration

Exponential increases in data acquisition rates necessitate new computational platforms and methods. Researchers today are already generating more data than we have the computational capacity to analyze. By 2025, an estimated 2 to 40 exabytes (millions of terabytes) will be needed just to store human genomes.[43] Beyond the mere cost considerations, current models cannot integrate data at this scale, requiring both advanced storage abilities and novel algorithmic development.[44] While this problem is most acute for unprocessed data, such as raw read or mass spectrometry files, scientists will eventually generate enough processed data to challenge the capacity of current data analysis methods.

Machine learning (ML) models have become an increasingly popular tool for analyzing data, despite their need for advanced computational resources. ML models thrive when given (1) large, extensive, curated training data; (2) complex architectures enabling model expressivity (ie, the ability of a model to estimate increasingly complex functions); and (3) the requisite hardware to fit these models. State-of-the-art models now require substantially more training and use hundreds or thousands of graphical processing units (GPUs) at a significant cost (compute resource/money/energy). ML scientists focused on advancing capabilities have largely ignored these costs, but identifying BASs will require more resource-efficient models that can yield high precision and recall in analyzing continuous streams of biological data.

Developing a BAS is, by itself, insufficient to meaningfully improve the state of public health and biodefense. Operationalizing a BAS, however, requires additional work that often involves adapting methods to reflect the realities of data acquisition, movement, and processing in real-world settings. For example, methods developed in a research setting generally have simultaneous access to all the data generated for a study. Such methods require adaptation for situations in which data are continuously generated in real time, such as during clinical biosurveillance or environmental sampling. Computational methods that can continuously analyze streaming data will, therefore, need to be developed. In addition, these methods must be computationally efficient and ideally require minimal computational resources. This is not yet possible with the current implementation of statistical models.

Improved Omics Analyses for Systems Biology Are Necessary

Standard Analyses of Omics Data Are Not Agent Agnostic

Although omics technologies, such as genomics and proteomics, have revolutionized our ability to detect pathogens and the immunological responses to them, the data that result from these technologies are typically analyzed in an agent-specific manner. For example, short-read sequencing is currently one of the most dominant genomic technologies used in the public health and biodefense fields, and the resulting reads are almost always aligned against reference genomes.

Without reference genomes, we must assemble the sequences de novo, which is a formidable task given the standard short reads from sequencers; our limitation for de novo assembly is exponentially increased for environmental samples. As a result, pathogen detection in environmental samples is performed using PCR. PCR is not an agent agnostic assay, however, because primers (ie, known targets) need to be constructed for use during amplification.

While we have broadly discussed the computational challenges associated with BAS development, the challenges and opportunities for BAS development will differ across sample type. For example, samples obtained from environmental monitoring, such as wastewater, are more complex and computationally difficult to analyze than clinical samples. Therefore, scalable computational models will likely be required sooner for environmental samples than for clinical samples.

Analysis of data from other omics technologies, such as proteomics and metabolomics, suffers from the same challenge. The standard proteomics data analysis involves detecting peptides in a sample by comparing experimentally derived spectra against a database of theoretically derived spectra.[45] These theoretically derived spectra are generated from peptide sequences, which are themselves derived from a reference genome. If one knows the identity of a sample, then a reference-based approach yields better results. In the case of metabolomics, untargeted metabolite identification usually requires the use of spectral libraries, which are typically generated from analyzing synthesized chemical standards.[46,47]

Continued Development and Adoption of De Novo Methods Are Needed

Creating BASs will require developing and adopting de novo algorithms for extremely complex samples. For example, long-read sequencing has the potential to greatly facilitate de novo genome and transcriptome assembly of pathogens and hosts. Although assembling genomes solely on short-sequence reads is generally feasible,[48] it fails on repetitive regions[49] and complex metagenomic datasets.[50,51] Long-read sequencing technologies overcome these challenges, resulting in the completion of the human reference genome[52] and low-complexity microbiomes.[53] These technologies have been embraced by the academic community, but they have yet to be widely used in the biodefense or public health spheres. Improvements to decrease error rates[54] are still needed, as is fundamental research into characteristic differences between long-read transcriptomics data and long-read genomics data.[54] Furthermore, additional development is required in order to allow usage in complex microbial environmental monitoring.

Similarly, advances in de novo analysis of proteomics data are necessary before integration in public health or biodefense. Although de novo peptide detection algorithms exist,[55,56] additional work is needed to increase analytic power, increase speed of data collection and analysis, decrease error rates, and increase utility with respect to reference-based methods.[57] In addition, challenges remain with detection and localization of novel and rare post-translational modifications (PTMs).[58] The field has almost exclusively focused on phosphoproteomics at the expense of other PTMs, such as glycosylation, which poses a particular challenge due to the diversity of its complex structures.[59] The ability to identify glycosylated biomolecules is important because they have been strongly implicated in host–pathogen interactions[60,61] and because 60% of proteins are possibly glycosylated.[62] Finally, as new nonmass spectrometry-based technologies are developed, such as single-molecule protein sequencing,[62-64] new computational methods will be needed to analyze the resulting data with their unique biases, assumptions, strengths, and weaknesses, as well as to integrate the data with that of other technologies, as described above.

The use of metabolomics in identifying BASs suffers from several computational challenges. The majority of untargeted metabolite identification uses spectral library searching of liquid chromatography-tandem mass spectrometry data.[46,47] This method is insufficient because it requires previous analysis of an extensive set of chemical standards, estimated to exceed the number of atoms in the universe.[65] This necessitates the improvement of current methods—which currently have poor performance and return numerous candidates—that can predict structures, spectra, and functions of completely novel chemicals.[66-68]

Improvements for Integrating Multimodal Data Are Needed

BAS development will likely rely on signatures that integrate multiple data sources to obtain a signal that is more robust to noise and error. Using multiple, different assays can also detect complementary signals that a single assay would be unable to detect. For example, genomics can describe an organism's potential processes but cannot inform which are active; conversely, proteomics can determine active processes but is unable to contextualize. Together, we obtain a more holistic understanding of the presence of a pathogen and the host response.

Developing multimodal signatures is not trivial,[69] requiring extensive interdisciplinary research and coordination among researchers and funding agencies. Common challenges include determining metadata sets and identifiers, and functional linking of different schemes. In addition, merging data from different assays is challenging because each assay has differing biases, variances, and power. For example, proteomics data acquired via data-dependent acquisition has missing data that correlates with abundance,[70] while transcriptomics data has missing data that correlates with time.[71] Furthermore, there are many different integration strategies,[72] necessitating generalized integration methods that can be applied to a large number of assays.

The recent growth of single-cell omics technology provides new tools enabling multimodal signatures,[21,73,74] particularly because such methods may mitigate the challenges of computationally integrating disparate datasets across assays. The technology is still immature, however, and realizing its potential will require investments to understand how biases may affect correlations between modes. Until we understand these effects, their utility is limited in the search for immunological BASs.

The interrogation and analysis of biological pathways is another promising avenue for generating BASs.[4,15] Fully understanding a pathway requires a systems biology approach that is most successful when multiple data streams, such as genomics, metabolomics, proteomics, and glycomics, are integrated for a single purpose (eg, immunomics). Multimodal analysis has the potential to generate signatures based on a small number of pathways related to pathogenicity and immune response against potential biothreats, especially if we can leverage single-cell technologies. Additional work is required to improve multimodal analysis before pathway-based approaches can be successful.

Final Thoughts

Although there is momentum to embrace bioagent agnostic biosurveillance, for instance by applying omics monitoring of wastewater,[75,76] the utility and impact of these technologies will be limited by the gaps in current computational and systems biology. Assays and technologies are starting to be developed to generate BASs, but computational analyses and tools need further investment, as delineated in this commentary. We advocate for converging technology, data, and modeling as a unified goal to enable BAS identification, characterization, and ultimately detection. From the computational side, the community must standardize data collection and annotation, increase research in scalable algorithms and methods development (especially in data harmonization for multimodal analysis), and implement tools whose computational throughput can match the levels attained by new technologies such as single-cell transcriptomics. Research in these disparate areas must be unified, and dedicated investment is needed to convert these basic research areas into actionable capabilities.

In this commentary, we have highlighted the human immune system as a potential sensor for a threat agnostic biosurveillance system, but other sample types can also potentially act as sensors. For example, interkingdom interactions between bacteria, archaea, and fungi in environmental samples may yield signals related to antibiotic resistance.[77] In addition, although we have focused on human pathogens, this approach is generalizable and can address threats to crops, livestock, and the environment on a larger scale.

One important aspect to note is that BASs will evolve, and so we emphasize the need to update these markers through a data-centric continual integration approach. Our computational models will need to be refactored as pathogens evolve. The evolution of pathogens underscores the need for bioagent agnostic signatures, which permits a flexibility in pattern recognition rather than focusing on a predetermined set of pathogens. While a BAS-based approach will not be the final solution, a biothreat agnostic approach will help close gaps that arise in list-based approaches.

Acknowledgments

Pacific Northwest National Laboratory is a multiprogram national laboratory operated by Battelle Memorial Institute for the US Department of Energy under contract DE-AC06-76RLO. The views and conclusions contained in this commentary are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US government. We would like to acknowledge the assistance of Crystal Devora, Aylin Sanchez, Miguel Gutierrez, and Luisa Rodriguez in generating data for this commentary. This work was funded in part by the National Institutes of Health award SC1AI148753.

References

1. Federal Select Agent Program. Select agents and toxins list. Updated August 1, 2023. Accessed December 19, 2023. https://www.selectagents.gov/sat/list.htm
2. World Health Organization. Prioritizing diseases for research and development in emergency contexts. Accessed December 19, 2023. https://www.who.int/activities/prioritizing-diseases-for-research-and-development-in-emergency-contexts

3. National Institute of Allergy and Infectious Diseases (NIAID). NIAID emerging infectious diseases/pathogens. Updated July 26, 2018. Accessed December 19, 2023. https://www.niaid.nih.gov/research/emerging-infectious-diseases-pathogens
4. Leiser OP, Hobbs EC, Sims AC, Korch GW, Taylor KL. Beyond the list: bioagent-agnostic signatures could enable a more flexible and resilient biodefense posture than an approach based on priority agent lists alone. *Pathogens.* 2021; 10(11):1497.
5. Mahroum N, Seida I, Esirgün SN, Bragazzi NL. The COVID-19 pandemic - How many times were we warned before? *Eur J Intern Med.* 2022;105:8-14.
6. Seligman L, Banco E. New worldwide threats prompt Pentagon to overhaul chem-bio defenses. *Politico.* January 9, 2023. Accessed December 19, 2023. https://www.politico.com/news/2023/01/09/russia-china-chemical-biological-weapons-pentagon-00077035
7. Kaslow RA, Stanberry LR, LeDuc JW, eds. *Viral Infections of Humans: Epidemiology and Control.* New York: Springer; 2014.
8. Peruski AH, Peruski LF Jr. Immunological methods for detection and identification of infectious disease and biological warfare agents. *Clin Diagn Lab Immunol.* 2003;10(4): 506-513.
9. Klebes A, Kittel AS, Verboket RD, von Stetten F, Früh SM. Multianalyte lateral flow immunoassay for simultaneous detection of protein-based inflammation biomarkers and pathogen DNA. *Sens Actuators B Chem.* 2022;355:131283.
10. Kang J, Jang H, Yeom G, Kim MG. Ultrasensitive detection platform of disease biomarkers based on recombinase polymerase amplification with H-sandwich aptamers. *Anal Chem.* 2021;93(2):992-1000.
11. Law JWF, Ab Mutalib NS, Chan KG, Lee LH. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. *Front Microbiol.* 2015;5:770.
12. Bernth Jensen JM, Petersen MS, Ellerman-Eriksen S, et al. Abundant human anti-Galα3Gal antibodies display broad pathogen reactivity. *Sci Rep.* 2020;10(1):4611.
13. Morse MA, Crosby EJ, Force J, et al. Clinical trials of self-replicating RNA-based cancer vaccines. *Cancer Gene Ther.* 2023;30(6):803-811.
14. Zhou J, Kaiser A, Ng C, et al. CD8+ T-cell mediated anti-malaria protection induced by malaria vaccines; assessment of hepatic CD8+ T cells by SCBC assay. *Hum Vaccin Immunother.* 2017;13(7):1625-1629.
15. Phillips SMB, Bergstrom C, Walker B, et al. Engineered cell line imaging assay differentiates pathogenic from non-pathogenic bacteria. *Pathogens.* 2022;11(2):209.
16. Tanner SD, Baranov VI, Ornatsky OI, Bandura DR, George TC. An introduction to mass cytometry: fundamentals and applications. *Cancer Immunol Immunother.* 2013;62(5):955-965.
17. Spitzer MH, Nolan GP. Mass cytometry: single cells, many features. *Cell.* 2016;165(4):780-791.
18. Zhou Y, Zhang Y, Johnson A, Venable A, Griswold J, Pappas D. Combined CD25, CD64, and CD69 biomarker panel for flow cytometry diagnosis of sepsis. *Talanta.* 2019; 191:216-221.
19. Chrdle A, Tinavská P, Dvořáčková O, et al. Early diagnosis of tularemia by flow cytometry, Czech Republic, 2003-2015[1]. *Emerg Infect Dis.* 2019;25(10):1919-1927.
20. Fulcher JM, Markillie LM, Mitchell HD, et al. Parallel measurement of transcriptomes and proteomes from same single cells using nanodroplet splitting. Preprint. *bioRxiv.* Posted May 18, 2022. Accessed December 19, 2023. doi:10.2172/1998955
21. Coates LC, FitzGerald O, Helliwell PS, Paul C. Psoriasis, psoriatic arthritis, and rheumatoid arthritis: is all inflammation the same? *Semin Arthritis Rheum.* 2016;46(3):291-304.
22. Furman D, Campisi J, Verdin E, et al. Chronic inflammation in the etiology of disease across the life span. *Nat Med.* 2019;25(12):1822-1832.
23. Hunter P. The inflammation theory of disease. The growing realization that chronic inflammation is crucial in many diseases opens new avenues for treatment. *EMBO Rep.* 2012; 13(11):968-970.
24. Shibru B, Fey K, Fricke S, et al. Detection of immune checkpoint receptors – a current challenge in clinical flow cytometry. *Front Immunol.* 2021;12:694055.
25. de Figueiredo AM, Glória JC, Chaves YO, Neves WLL, Mariúba LAM. Diagnostic applications of microsphere-based flow cytometry: a review. *Exp Biol Med (Maywood).* 2022;247(20):1852-1861.
26. Raman AT. A research parasite's perspective on establishing a baseline to avoid errors in secondary analyses. *Gigascience.* 2021;10(3):giab015.
27. Mogilenko DA, Shpynov O, Andhey PS, et al. Comprehensive profiling of an aging immune system reveals clonal GZMK$^+$ CD8$^+$ T cells as conserved hallmark of inflammaging. *Immunity.* 2021;54(1):99-115.e12.
28. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018. Published correction appears in *Sci Data*. 2019;6(1):6.
29. Hughes LD, Tsueng G, DiGiovanna J, et al; NIAID Systems Biology Data Dissemination Working Group. Addressing barriers in FAIR data practices for biomedical data. *Sci Data.* 2023;10(1):98.
30. Rothstein MA. Informed consent for secondary research under the new NIH Data Sharing policy. *J Law Med Ethics.* 2021;49(3):489-494.
31. Martens L, Hermjakob H, Jones P, et al. PRIDE: the proteomics identifications database. *Proteomics.* 2005;5(13): 3537-3545.
32. Vizcaíno JA, Csordas A, Del-Toro N, et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* 2016;44(22):11033.
33. Sze MA, Schloss PD. Looking for a signal in the noise: revisiting obesity and the microbiome. *mBio.* 2016;7(4): e01018-16. Published correction appears in *mBio*. 2017; 8(6):e01995-17.
34. Pilgrim J. The opportunities of research parasitism: a case study using the Barcode of Life Data System (BOLD). *Gigascience.* 2022;11:giac123.
35. Walzer M, García-Seisdedos D, Prakash A, et al. Implementing the reuse of public DIA proteomics datasets: from the PRIDE database to Expression Atlas. *Sci Data.* 2022; 9(1):335.

36. Won D, Kim SH, Kim B, Lee ST, Kang HC, Choi JR. Reanalysis of genomic sequencing results in a clinical laboratory: advantages and limitations. *Front Neurol.* 2020;11: 612.
37. Phua SX, Lim KP, Goh WWB. Perspectives for better batch effect correction in mass-spectrometry-based proteomics. *Comput Struct Biotechnol J.* 2022;20:4369-4375.
38. Čuklina J, Lee CH, Williams EG, et al. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Mol Syst Biol.* 2021;17(8):e10240.
39. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet.* 2010;11(10):733-739.
40. Tran HTN, Ang KS, Chevrier M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* 2020;21(1):12.
41. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 2010;11(10):R106.
42. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010;11(3):R25.
43. Stephens ZD, Lee SY, Faghri F, et al. Big data: astronomical or genomical? *PLoS Biol.* 2015;13(7):e1002195.
44. Muir P, Li S, Lou S, et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* 2016;17:53. Published correction appears in *Genome Biol.* 2016;17:78.
45. Eng JK, Searle BC, Clauser KR, Tabb DL. A face in the crowd: recognizing peptides through database search. *Mol Cell Proteomics.* 2011;10(11):R111.009522.
46. Bittremieux W, Wang M, Dorrestein PC. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics.* 2022;18(12):94.
47. Neumann S, Böcker S. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal Bioanal Chem.* 2010;398(7-8):2779-2788.
48. Paszkiewicz K, Studholme DJ. De novo assembly of short sequence reads. *Brief Bioinform.* 2010;11(5):457-472.
49. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet.* 2011;13(1):36-46. Published correction appears in *Nat Rev Genet.* 2012;13(2):146.
50. Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. *Brief Bioinform.* 2020;21(2):584-594.
51. Marx V. Method of the year: long-read sequencing. *Nat Methods.* 2023;20(1):6-11.
52. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science.* 2022;376(6588):44-53.
53. Somerville V, Lutz S, Schmid M, et al. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol.* 2019; 19(1):143.
54. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 2020;21(1):30.
55. Tran NH, Qiao R, Xin L, et al. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods.* 2019;16(1):63-66.
56. Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A.* 2017; 114(31):8247-8252.
57. Muth T, Renard BY. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinform.* 2018;19(5): 954-970.
58. Miller RM, Smith LM. Overview and considerations in bottom-up proteomics. *Analyst.* 2023;148(3):475-486.
59. Grabarics M, Lettow M, Kirschbaum C, Greis K, Manz C, Pagel K. Mass spectrometry-based techniques to elucidate the sugar code. *Chem Rev.* 2022;122(8):7840-7908.
60. Lin B, Qing X, Liao J, Zhuo K. Role of protein glycosylation in host-pathogen interaction. *Cells.* 2020;9(4):1022.
61. Qin R, Mahal LK. The host glycomic response to pathogens. *Curr Opin Struct Biol.* 2021;68:149-156.
62. Timp W, Timp G. Beyond mass spectrometry, the next step in proteomics. *Sci Adv.* 2020;6(2):eaax8978.
63. Yu L, Kang X, Li F, et al. Unidirectional single-file transport of full-length proteins through a nanopore. *Nat Biotechnol.* 2023;41(8):1130-1139. Published correction appears in *Nat Biotechnol.* 2023;41(10):1483.
64. Alfaro JA, Bohländer P, Dai M, et al. The emerging landscape of single-molecule protein sequencing technologies. *Nat Methods.* 2021;18(6):604-617.
65. Lemonick S. Exploring chemical space: Can AI take us where no human has gone before? *C&EN.* 2020;98(13).
66. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A.* 2015; 112(41):12580-12585.
67. Stravs MA, Dührkop K, Böcker S, Zamboni N. MSNovelist: de novo structure generation from mass spectra. *Nat Methods.* 2022;19(7):865-870.
68. van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci U S A.* 2016;113(48):13738-13743.
69. Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat Comput Sci.* 2021;1(6):395-402.
70. Matafora V, Bachi A. Missing value monitoring to address missing values in quantitative proteomics. *Methods Mol Biol.* 2021;2228:401-408.
71. Chan LY, Mugler CF, Heinrich S, Vallotton P, Weis K. Non-invasive measurement of mRNA decay reveals translation initiation as the major determinant of mRNA stability. *eLife.* 2018;7:e32536.
72. Canzler S, Schor J, Busch W, et al. Prospects and challenges of multi-omics data integration in toxicology. *Arch Toxicol.* 2020;94(2):371-388.
73. Swanson E, Lord C, Reading J, et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife.* 2021;10: e63632.
74. Nam AS, Kim KT, Chaligne R, et al. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature.* 2019;571(7765):355-360.
75. Cui L, Li HZ, Yang K, Zhu LJ, Xu F, Zhu YG. Raman biosensor and molecular tools for integrated monitoring of

pathogens and antimicrobial resistance in wastewater. *Trends Analyt Chem.* 2021;143:116415.

76. Kumaraswamy R, Amha YM, Anwar MZ, Henschel A, Rodríguez J, Ahmad F. Molecular analysis for screening human bacterial pathogens in municipal wastewater treatment and reuse. *Environ Sci Technol.* 2014;48(19):11610-11619.
77. Schmidt R, Cordovez V, de Boer W, Raaijmakers J, Garbeva P. Volatile affairs in microbial interactions. *ISME J.* 2015; 9(11):2329-2335.

Address correspondence to:
Tony Chiang
Pacific Northwest National Laboratory
1100 Dexter Ave N
Suite 500
Seattle, WA 98109

Email: tony.chiang@pnnl.gov