How large biomolecular and clinical datasets unite

Yves A. Lussier, MD
Center for Biomedical Informatics & Biostatistics, Department of Medicine,
BIO5 Institute, Cancer Center, The University of Arizona
1230 North Cherry Avenue, Tucson, AZ 85721
Yves@email.arizona.edu

Atul J. Butte, MD, PhD
Bakar Computational Health Sciences Institute,
University of California, San Francisco
550 16th Street, San Francisco, CA 94158
Atul.Butte@ucsf.edu

Haiquan Li, PhD
College of Agriculture and Life Sciences, The University of Arizona
1177 E 4th Street, Shantz 509, Tucson, AZ 85719
Haiquan@email.arizona.edu

Rong Chen, PhD
Sema4 Genomics
333 Ludlow Street, Stamford, CT 06902
Rong.Chen@seme4Genomics.com

Jason H. Moore, PhD
The Perelman School of Medicine, University of Pennsylvania
D202 Richards, 3700 Hamilton Walk, Philadelphia, PA 19104
Jhmoore@upenn.edu

This paper summarizes the workshop content on how the integration of large biomolecular and clinical datasets can enhance the field of population health via translational informatics. Large volumes of data present diverse challenges for existing informatics technology, in terms of computational efficiency, modeling effectiveness, statistical computing, discovery algorithms, and heterogeneous data integration. While accumulating large 'omics measurements on subjects linked with their electronic record remains a challenge, this workshop focuses on non-trivial linkages between large clinical and biomolecular datasets. For example, exposures and clinical datasets can relate through zip codes, while comorbidities and shared molecular mechanisms can relate diseases. Workshop presenters will discuss various methods developed in their respective labs/organizations to overcome the difficulties of combining together such large complex datasets and knowledge to enable the translation to clinical practice for improving health outcomes.

Keywords: Translational informatics, biomolecular, clinical, population health, big data, workshop

1. Introduction, Background, and Motivation

The field of population health is rapidly moving to the forefront of research, with the advancement of biotechnologies and growth of international collaborations enabling the vast

accumulation of population health data. The availability of such data crossing multiple dimensions, from electronic health records, lifestyles, environmental factors, genetics, to genomics, is promising for further advancing the field via translational bioinformatics. A growing trend is the integrative data collection that encompasses all aspects (both genetic and non-genetic factors) of the same participants, exemplified by eMERGE[1], UK Biobank[2], and All of US program[3], among many others in specific domain and specialties.[4,5]

However, large volumes of data present diverse challenges for existing informatics technology, in terms of computational efficiency, modeling effectiveness, statistical computing, discovery algorithms, and heterogeneous data integration. These new demands also call for bridging the gap between disciplines among statistical genetics, health informatics, and bioinformatics. Successful endeavors in these areas will dramatically enhance the understanding of the genetic/epigenetic mechanisms of complex diseases and their interplay with the environment and lifestyles as well as foster the translation of these findings to clinical practice to improve health outcomes.[6,7]

In this era of Big Data science, the number of opportunities to study large-scale molecular and population datasets together is flourishing. The developers of PheWAS[8] were among the pioneers to transform heterogeneous and sparsely-annotated clinical data for systematic analysis with densely-annotated SNP arrays. Combining this knowledge with publicly-available data from sources, such as UK Biobank[2] that offers health information on over 500,000 participants, not only promotes Big Data analytics, but also demonstrates the feasibility of such studies.

This paper summarizes how the integration of large datasets, such as biomolecular and clinical data, can advance the field of population health via translational informatics as well as focuses on current approaches to overcome the challenges of combining these complex data.

2. Workshop Presenters

The three-hour workshop is organized in the form of six presentations, including two keynote speakers, followed by a discussion session, which will be moderated by Dr. Yves A. Lussier.

Keynote speakers are:

- Atul Butte, MD, PhD (Priscilla Chan and Mark Zuckerberg Distinguished Professor, University of California, San Francisco)
- Jason H. Moore, PhD, FACMI (Edward Rose Professor of Informatics, University of Pennsylvania)

Additional speakers include:

- Francesca Vitali, PhD (Research Assistant Professor, The University of Arizona)
- Lara M. Mangravite, PhD (President, Sage Bionetworks)
- Serghei Mangul, PhD (QCB Postdoctoral Fellow, University of California, Los Angeles)
- Marina Sirota, PhD (Assistant Professor, University of California, San Francisco)

3. Presenters' Abstracts

Translating a trillion points of data into therapies, diagnostics, and new precision medicine
Atul Butte, MD, PhD (University of California, San Francisco)
There is an urgent need to take what we have learned in our new "genome era" and use it to create a new system of precision medicine, delivering the best preventative or therapeutic intervention at the right time, for the right patients. Dr. Butte's lab at the University of California, San Francisco builds and applies tools that convert trillions of points of molecular, clinical, and epidemiological data -- measured by researchers and clinicians over the past decade and now commonly termed "big data" -- into diagnostics, therapeutics, and new insights into disease. Several of these methods or findings have been spun out into new biotechnology companies. Dr. Butte, a computer scientist and pediatrician, will highlight his lab's recent work, including the use of publicly-available molecular measurements to find new uses for drugs including new therapies for autoimmune diseases and cancer, discovering new druggable targets in disease, the evaluation of patients and populations presenting with whole genomes sequenced, integrating and reusing the clinical and genomic data that result from clinical trials, discovering new diagnostics include blood tests for complications during pregnancy, and how the next generation of biotech companies might even start in your garage.

Enabling translational bioinformatics with accessible artificial intelligence
Jason H. Moore, PhD (University of Pennsylvania)
Artificial intelligence (AI) is a rapidly maturing technology that has the potential to accelerate translational bioinformatics and precision medicine using both basic science and clinical data. While AI has become widespread, many commercial AI systems are not yet accessible to individual researchers nor the general public due to the deep knowledge of the systems required to use them. We believe that AI has matured to the point where it should be an accessible technology for everyone. We present an ongoing project whose goal is to deliver an open-source, user-friendly AI system that is specialized for machine learning analysis of complex data in the biomedical and health care domains.

Novel and emerging data fusion strategies for integrating health and biomolecular data
Francesca Vitali, PhD (The University of Arizona)
Over the last few years, biomedical research and clinical practice have experienced incredible growth in terms of both the amount and variety of data being collected and leveraged for different types of analysis. This represents a great opportunity to increase our knowledge about many biological mechanisms as well as improve the medical process. However, not all big data is created equal, complicating the integration and analysis of such large datasets. For example, clinical record data is highly heterogeneous, sparsely annotated, and contains several measurement types and unstructured text fields comprised of ambiguous statements as well as varying levels of certainty, whereas genomic and imaging data are crisp, homogeneous, densely annotated data with a low cardinality of distinct variables. Nowadays, the development of novel methodologies capable of integrating population health data with biomolecular data is crucial,

not only for enabling translational and clinical research, but for developing more effective patient care. However, integrating these data are particularly challenging when the molecular measurements are not conducted on individual subjects. In order to take full advantage of the wide spectrum of biomedical data available, advanced data integration tools need to be developed. In this context, we will discuss novel and emerging data fusion strategies for integrating health and biomolecular data to develop new research hypotheses and conduct predictive and data interpolation operations. These methods include approaches that (i) take into account comprehensive drug-exposure histories of individuals derived from healthcare data, while also including genetic, environmental, and lifestyle variabilities for each individual; (ii) integrate electronic medical records with biobank data to identify new disease pathways; (iii) combine multi-omic profiling with clinical factors from large cohorts; and (iv) perform crisp integration of bimolecular data whilst leveraging population measurements (e.g., counties, medication, diseases).

Open practices to advance biomedicine through data-intensive science

Lara M Mangravite, PhD (Sage Bionetworks)

Open science practices in bio-computing have been promoted over the past 10 years under the premise that these approaches can improve confidence and, therefore, speed advancement of biomedical hypotheses stemming from computational research. In that time, we have observed wide adoption of open practices including those focused on open data, open commons(es), open source software, and open access publishing. Although many of these efforts help to establish confidence in research observations amongst computationally-savvy researchers, they often fail to support the wider acceptance necessary to inform trajectories of biological inquiry and/or to promote adoption for use in clinical care. Here, we discuss complementary mechanisms to further support the advancement of biocomputational hypotheses, including those developed using emerging digital health technologies, through the transfer and translation of knowledge across research domains.

Seeing Beyond the Target: Constructing germline research cohorts from clinical tumor sequencing

Serghei Mangul, PhD (University of California, Los Angeles)

Tens of thousands of cancer patients have had their tumors sequenced to identify clinically actionable mutations. In addition to saving lives, this activity has produced valuable research data sets leading to significant discoveries in basic and translational domains. However, the targeted nature of clinical tumor sequencing has a limited research scope, especially with respect to germline genetics. In this work, we address this problem by developing a software platform (SBT: Seeing Beyond the Target) that mines discarded tumor sequences to produce rich research level data including genome-wide germline genotypes, T and B cell receptor sequences, rDNA and mtDNA copy number, and HLA types. These features have been demonstrated as potential prognostic indicators in research studies, and our methods now make them available in large-scale clinical cohorts. We validate the accuracy of our tool, by comparison, to deeply sequenced cohorts and show its utility through replication of known genetic associations. We provide a free downloadable cloud implementation and demonstrate its efficiency by constructing the largest

germline-somatic cohort produced to date (n>20,000), more than doubling the size of The Cancer Genome Atlas. We believe that SBT will greatly increase the research potential of clinical tumor data sets and provide a bridge between the germline and somatic research communities. SBT is freely available at https://github.com/smangul1/seeing.beyond.target/wiki

Leveraging population level molecular, environmental and clinical data to study adverse pregnancy outcomes

Marina Sirota, PhD (University of California, San Francisco)

Given the wealth and availability of genomic, clinical and environmental exposure data, computational integrative methods provide a powerful opportunity to identify population-specific determinants of disease. In this talk, I will discuss our efforts to develop computational methods and integrate large-scale genomic, transcriptomic and environmental exposure datasets to elucidate factors that affect preterm birth (PTB). Preterm birth, or the delivery of an infant prior to 37 weeks of gestation, is a major health concern. Infants born prematurely, comprising of about 12% of the US newborns, have elevated risks of neonatal mortality and a wide array of health problems. In our work, we leverage the rich multi-omic, clinical and environmental variation data to advance our understanding of biology of preterm birth as it relates to all populations. Our findings further inform precise population-specific diagnostic and therapeutic strategies bringing us closer to applying precision medicine to this important biomedical problem.

4. Conclusion

This workshop will highlight a number of methods, strategies, and tools currently being developed for integrating population health data with biomolecular data to mitigate the diverse challenges of existing informatics technology. The ability to combine these big data across various domains to conduct meaningful and interpretable analysis is critical for improving overall population health outcomes.

Acknowledgements

We would like to thank Dr. Colleen Kenost for her organizational contributions with this workshop and proceedings paper.

References

1. O. Gottesman, H. Kuivaniemi, et al., *Genetics In Medicine*, 2013, **15**, 761.
2. R. Collins, *The Lancet*, 2012, **379**, 1173-1174.
3. N. I. o. Health, 2018.
4. R. J. Hodes and N. Buckholtz, *Expert Opinion on Therapeutic Targets*, 2016, **20**, 389-391.
5. L. National Heart, and Blood Institute, 2016.
6. H. Li, I. Achour, et al., *Npj Genomic Medicine*, 2016, **1**, 16006.
7. L. Li, W.-Y. Cheng, et al., *Science Translational Medicine*, 2015, **7**, 311ra174-311ra174.
8. J. C. Denny, L. Bastarache, et al., *Nature Biotechnology*, 2013, **31**, 1102.