Pacific Symposium on Biocomputing 25:743-747(2020)

# Translational Bioinformatics: Biobanks in the Precision Medicine Era

Marylyn D Ritchie
*Department of Genetics and Institute for Biomedical Informatics, The Perelman School of Medicine, University of Pennsylvania, A301 Richards Building, 3700 Hamilton Walk*
*Philadelphia, PA 19104, USA*
*Email: marylyn@pennmedicine.upenn.edu*

Jason H Moore
*Department of Biostatistics, Epidemiology, & Informatics, Department of Genetics, and Institute for Biomedical Informatics, The Perelman School of Medicine, University of Pennsylvania, D202 Richards Building, 3700 Hamilton Walk*
*Philadelphia, PA 19104, USA*
*Email: jhmoore@upenn.edu*

Ju Han Kim
*Department of Biomedical Sciences, Seoul National University Graduate School, Biomedical Science Building 117,*
*103 Daehakro, Jongro-gu, Seoul 110-799, Korea*
*Email: juhan@snu.ac.kr*

Translational bioinformatics (TBI) is focused on the integration of biomedical data science and informatics. This combination is extremely powerful for scientific discovery as well as translation into clinical practice. Several topics where TBI research is at the leading edge are 1) the use of large-scale biobanks linked to electronic health records, 2) pharmacogenomics, and 3) artificial intelligence and machine learning. This perspective discusses these three topics and points to the important elements for driving precision medicine into the future.

*Keywords:* translational bioinformatics, precision medicine, pharmacogenomics, artificial intelligence, machine learning, electronic health records, biobank

## 1. Introduction

Translational bioinformatics (TBI) is a multi-disciplinary and rapidly emerging field of biomedical data sciences and informatics that includes the development of technologies that efficiently translate basic molecular, genetic, cellular, and clinical data into clinical products or health implications. TBI is a relatively young discipline that spans a wide spectrum from big data to comprehensive analytics to diagnostics and therapeutics. TBI involves applying novel methods to the storage, analysis, and interpretation of a massive volume of genetics, genomics, multi-omics, and clinical data; this includes diagnoses, medications, laboratory measurements, imaging, and

© 2019 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

clinical notes. TBI bridges the gap between bench research and real-world applications to human health. Many health-related topics are increasingly falling within the scope of TBI, including rare and complex human disease, cancer, biomarkers, pharmacogenomics, drug repositioning, genomic medicine, and clinical decision support systems.

TBI in precision medicine attempts to determine individual solutions based on the genomic, environmental, and clinical profiles of each individual, providing an opportunity to incorporate individual genomic data into patient care. While a plethora of genomic signatures have successfully demonstrated their predictive power, they are merely statistically significant differences between dichotomized phenotypes (for example cases and controls of a specific disease) that are in fact severely heterogeneous phenotypes. Despite many translational barriers, connecting the molecular world to the clinical world and vice versa will undoubtedly benefit human health in the near future.

Due to the rapid pace of TBI, we assembled diverse perspectives to review the state of the art in translation bioinformatics using biobanks in this era of precision medicine. Discovery and implementation research projects are underway in academic medical centers, biotech companies, pharmaceutical companies, and health systems. We provide perspective on where the current efforts are focused and where the future is headed for biobanks in these different disciplines. Special attention will be given to pharmacogenomics, which is how individuals respond to therapy based on genetic variation. Pharmacogenomics is one of the primary areas where both discovery and implementation are well underway across the globe. We also discuss artificial intelligence and machine learning and how these are being used now in biobanks as well as how we anticipate they will be used in the future. Translational bioinformatics is a fast-moving field and we believe that integrating the basic science community from genomics, bioinformatics, computer science, and statistics together with the translational community including clinical/medical informatics, pharmacogenomics, and genomic medicine will be mutually beneficial to accelerate the translational of biomedical research into precision medicine.

## 2. Biobanks linked to clinical data sources

In recent years, we have witnessed an explosion of data in biomedical sciences. Between the development of new high-throughput molecular technologies and the adoption of electronic health records (EHRs), the amount of data currently available for biomedical research is unprecedented. Electronic health records (EHRs) were originally developed in the 1960s, but became more pervasive in the mid-1990s[1]. Widespread adoption of EHRs in the United States took place after the American Recovery and Reinvestment Act mandated that all public and private healthcare providers were required to adopt them by January 1, 2014[2]. This led to the implementation of EHRs in healthcare centers around the country for the practice of medicine.

EHRs for biomedical research has only recently gained significant traction[3]. Moreover, around 2010, we also learned that these clinical data could be made useful specifically for genetic association analysis by using the clinical data integrated with genomic data from a biobank linked to the EHR[4–6]. The data collected to record patient medical conditions and for billing insurance companies captured the relevant data needed to define many phenotypes or diseases in research

participants[6,8]. For some traits, these EHR data can mimic similar data collections used in clinical trials, epidemiological studies, or national patient registries; though some very specific phenotypes may not be ascertained with the same level of specificity as in a clinical trial or epidemiological survey[8]. As such, careful consideration of the type of clinical data needed for a study is warranted when considering the use of EHR (or real world) data rather than clinical trials or epidemiologic study data. Similarly, population-based registries capture much of the same types of clinical data, albeit some are more focused around specific diseases or conditions and also more standardized and structured. These types of registries have been systematically collecting data for decades. Many research programs have capitalized on these population-based registries with complementary biobanks for research linkage to the health registry including deCODE[9], Genome Netherlands (GoNL)[10], and UK Biobank[11]. EHRs and national health registries have both been adopted as clinical data sources for genetic and genomic analyses for a wide variety of diseases/conditions.

The utility of these clinical data linked with genetic and genomic data has enormous potential for disease gene discovery. Much research is ongoing to identify risk factors for complex disease, evaluate the potential repurposing medications for multiple phenotypes, and the identification of novel therapeutic targets. In addition, the development of polygenic risk scores (PRS) as well as genomic risk assessments, which integrate PRS with known clinical risk factors, are an emerging area of research in large scale biobanks linked with clinical data sources. As more health systems and academic medical centers continue to build large scale biobanks, in addition to the growing industry of direct-to-consumer and recreational genomics, the opportunities for discovery in biobanks linked to clinical data sources will continue to explode.

## 3. Pharmacogenomics – discoveries for precision medicine

Substantial knowledge in pharmacogenomics (PGx) has been accumulated through genetic studies in human populations and model system datasets over the past two decades. Variability in drug response has often been found to be related to functional changes in genes/proteins due to inherited interindividual variation. The majority of PGx variation is present in genes important for drug absorption, distribution, metabolism, and/or excretion (also referred to as ADME genes). In addition to the ADME genes, other genes are involved in the pharmacokinetic (PK) and pharmacodynamic (PD) processing of the drugs are also important for drug response. Genetic variation in these ADME or PK/PD genes/proteins may greatly influence therapeutic outcomes such as drug efficacy and safety. Personalized drug therapy, where genetic information is considered in treatment planning, is especially advantageous for drugs with a narrow therapeutic window or when drug toxicity is serious or life-threatening. In many large-scale biobanks linked to an EHR, research focused on implementation or evaluating these genetic variants prior to treatment, is at the leading edge. This research is focused on the premise that safe and effective treatment decisions can be made at treatment initiation, reducing trial and error, if the genetic variant information is available prior to prescribing and availability to the clinician. Much ongoing research in both discovery of new PGx variants and implementation of known PGx variants takes place in large scale biobanks linked to EHRs. The beauty of the EHR is that researchers can collect information about medication prescribing, prescription fills, medication changes, and treatment outcomes. These datasets enable

multiple gene-drug experiments to be conducted in a single large-scale data set. We anticipate that the identification of addition genetic variants important for drug treatment response will emerge as more large-scale biobanks covering diverse ethnic groups linked to EHRs are developed. We also anticipate that implementation of PGx into clinics will accelerate in upcoming years. With this knowledge, prevention of serious adverse drug reactions can be prioritized and this will help to fully realize the potential of precision medicine.

## 4. Artificial intelligence and machine learning in biobanks

The integration of genomics data with electronic health record data opens the door to numerous research question about the role of genomic variation in human health. Artificial intelligence and machine learning have an important role to play in answering these questions. An important challenge that computational methods are well-suited to is the definition of phenotypes that are more accurate than those provided by disease diagnoses captured in billing codes. The challenge here to find a mathematical function of laboratory measures, medication, and other information that can be used to make a more accurate diagnosis. Machine learning is ideally suited to building models of disease phenotypes. Once accurate phenotypes are derived, the next step is to perform association analysis. Genome-wide association studies in epidemiologic studies have focused almost exclusively on statistical tests of each genetic variant independent of their genomic or environmental context. This has benefits such as speed and interpretation. However, genetic variants are likely to have effects that are context-dependent and thus not captured by univariate models. Machine learning can complement statistical methods by modeling non-additive effects among multiple factors. Further, machine learning can capture heterogeneity of genetic effects that can also be quite common. The development and application of machine learning methods in biobanks is an active area of research and very much in its infancy. Issues such as choosing the right machine learning methods for the data, interpreting the results, and developing actionable validation and implementation strategies are complex and in need of future work. An emerging area addresses the first issue is automated machine learning (AutoML) that focuses on optimization algorithms for choosing the right methods for a given data set. Automated machine learning is a step towards artificial intelligence with the goal of developing algorithms that solve problems the way human analysts do. It is important to remember that the goal of machine learning is to identify those unexpected results that would be missed by parametric statistical methods.

## 5. Discussion

Translational bioinformatics (TBI) lives at the intersection of informatics and biomedical data science. Due to the explosion of data in molecular and cellular technologies in the ‘omics era paired with the rapid increase in the access and availability to clinical information from electronic health records, the possibilities for discovery and rapid translational into clinically and biologically meaningful outcomes are tremendous. To all of these rich data, add the powerful technologies being developed in artificial intelligence and machine learning, this leads to a unique opportunity for biomedical data science to elevate in ways that are unprecedented. The future of precision medicine will be led by translational bioinformatics.

## References


1. Doyle-Lindrud, S. The evolution of the electronic health record. *Clin. J. Oncol. Nurs.* **19**, 153–154 (2015).
2. Federal Mandate for Electronic Medical Records. *USF Health Online* (2017). Available at: https://www.usfhealthonline.com/resources/healthcare/electronic-medical-records-mandate/. (Accessed: 20th November 2017)
3. Häyrinen, K., Saranto, K. & Nykänen, P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *Int. J. Med. Inf.* **77**, 291–304 (2008).
4. Ritchie, M. D. *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* **86**, 560–572 (2010).
5. Kho, A. N. *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Sci. Transl. Med.* **3**, 79re1 (2011).
6. Wilke, R. A. *et al.* The emerging role of electronic medical records in pharmacogenomics. *Clin. Pharmacol. Ther.* **89**, 379–386 (2011).
7. Roque, F. S. *et al.* Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.* **7**, e1002141 (2011).
8. Casey, J. A., Schwartz, B. S., Stewart, W. F. & Adler, N. E. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu. Rev. Public Health* **37**, 61–81 (2016).
9. deCODE genetics | a global leader in human genetics. Available at: https://www.decode.com/. (Accessed: 18th January 2018)
10. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22**, 221–227 (2014).
11. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).