Translational Bioinformatics: Integrating Electronic Health Record and Omics Data

Dokyoon Kim
Department of Biostatistics, Epidemiology, & Informatics, Institute for Biomedical Informatics,
The Perelman School of Medicine, University of Pennsylvania,
D202 Richards Building, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: dokyoon.kim@pennmedicine.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

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

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 clinical utility of polygenic risk scores, 2) data integration, 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, data integration, artificial intelligence, machine learning, electronic health records, biobank, polygenic risk scores

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 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 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 including the clinical utility of polygenic risk scores, data integration, and artificial intelligence in medicine. We provide perspective on where the current efforts are focused and where the future is headed for biobanks in different disciplines, especially about the utility of polygenic risk scores. Additionally, special attention will be given to data integration. In particular, radiogenomics or imaging genomics is one of the primary areas that focus on the relationship between imaging phenotypes and genomics. We also discuss artificial intelligence and machine learning and how these are being used now for integrating electronic health record (EHR) and omics data 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. The clinical utility of polygenic risk scores

Many research programs have capitalized on these population-based registries with complementary biobanks for research linkage to the health registry including UK Biobank [1], FinnGEN [2], and deCODE [3]. 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 particular, 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. PRS is a value accumulated based on the effect sizes of multiple genetic variants across the genome and has shown great promise in the prediction of risk for many diseases [4]. Furthermore, recent studies for many diseases suggest that our knowledge of the common variants underlying diseases or phenotypes has improved to a point where polygenic risk profiling provides personal and clinical utility by identifying groups of individuals who could benefit from the knowledge of their probabilistic susceptibility to disease [5]. As more health systems and academic medical centers continue to build large scale biobanks, the opportunities for discovery in biobanks linked to clinical data sources will continue to explode.

3. Data integration

While individual analysis of omics datasets is valuable for identifying omic-phenotype associations, analyses using only one data type are not sufficient to fully elucidate complex diseases because such diseases are the end point of events cumulating with multiple variations through multi-omics biology. To better understand the genetic architecture of complex diseases, relevant strategies for integrating multi-omics data are required. Many studies have shown that an integrative systems genomics approach and addressed the idea that integration of multi-omics data can be substantially more informative than separate analyses of each single dimension of genomic data [6]. Data integration methods can be broadly categorized into two types of approaches, as follows. In multi-staged analysis, models are constructed using only two different scales at a time, in a stepwise, linear, or hierarchical manner. A multi-staged analysis would be applicable when the relationship between genotype and phenotype can be modelled in a linear manner (e.g. association of SNPs with DNA methylation) and subsequently associated with phenotypes. However, this approach is difficult to apply simultaneously to more than two types of -omics data. An alternative approach is meta-dimensional analysis (i.e. fusion of scales), which simultaneously combines all scales of data to produce complex, meta-dimensional models with multiple variables from different data types. The scale and richness of these ever-increasing data sets hold great promise, yet the complexity presents an urgent need to find effective ways to integrate diverse data from different levels of technologies to fully exploit the potential informativeness of big data. One particularly rich source of information contained in medical records are imaging data, such as MRI, CT scan, fundoscopic images, or histopathology slides. Radiogenomics or imaging genomics is one of the primary areas that focus on the relationship between imaging phenotypes and genomics. With state-of-the-art deep learning approaches, radiogenomics might offer a practical way to leverage limited and incomplete data to generate knowledge that could lead to improved decision making, and as a result, improved patient outcomes [7].

4. Artificial intelligence in medicine

The integration of genomics data with EHR 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 and imaging data from EHRs, 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] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J: The UK Biobank resource with deep phenotyping and genomic data. Nature 2018, 562:203-9.

[2] Mars N, Koskela JT, Ripatti P, Kiiskinen TTJ, Havulinna AS, Lindbohm JV, Ahola-Olli A, Kurki M, Karjalainen J, Palta P, FinnGen, Neale BM, Daly M, Salomaa V, Palotie A, Widen E, Ripatti S: Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nature medicine 2020, 26:549-57.

[3] Swede H, Stone CL, Norwood AR: National population-based biobanks for genetic research. Genetics in medicine : official journal of the American College of Medical Genetics 2007, 9:141-9.

[4] Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S: Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 2018, 50:1219-24.

[5] Torkamani A, Wineinger NE, Topol EJ: The personal and clinical utility of polygenic risk scores. Nature reviews Genetics 2018, 19:581-90.

[6] Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D: Methods of integrating data to uncover genotype-phenotype interactions. Nature reviews Genetics 2015, 16:85-97.

[7] Mazurowski MA: Radiogenomics: what it is and why it is important. J Am Coll Radiol 2015, 12:862-6.