OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of
Data Mining in
Genomics & Proteomics

Perspective

Integrating Genomics Improves Large-Scale of Human Genomics

Franco Ferrari[*]

Department of Biotechnology, Illinois University, Chicago, USA

ABOUT THE STUDY

Because of larger cohorts and the development of more powerful prediction algorithms, Polygenic Risk Scores (PRSs) are becoming more relevant to public health. Accurate PRS predictors are now being trained to predict a variety of human diseases, including Type 2 diabetes, coronary artery disease, and breast cancer. Such PRS predictors are expected to become pervasive in clinical human health and decision-making, thus playing a critical role in the realization of personalized medicine. PRS predictors are classified into two types based on the type of training data used: those that use summary statistics from Genome-Wide Association Studies (GWAS) and those that use individual-level data. Because of larger sample sizes, the combined GWAS approach is more common today. However, this is rapidly changing as the size of individual-level human genetic variation data increases, with cohorts containing hundreds of thousands, if not millions, of people. Large individual-level cohorts are increasingly providing the opportunity to train accurate predictors for estimating PRSs that can outperform the combined GWAS approach. There are many well-established methods for training predictors on summary statistics and individual-level data today, but these predictors mostly investigate linear relationships. AI and Deep Learning (DL) have revolutionized several scientific fields and are changing our society. It was named one of the ten scientific events that shaped the last decade at the end of 2019. DL has gained traction in the life sciences in recent years. Imaging in particular, but also single-cell sequencing, protein localization, and protein folding. In the case of genome sequence data, efforts have primarily focused on identifying motifs, such as ChIP-seq, or identifying genomic variation. Simultaneously, DL frameworks for large discrete data sets, such as genome-wide data, have received little attention in the field. One potential advantage of DL-based PRS prediction methods is their ability to capture complex non-linear effects such as epistasis. Previous work using Neural Networks (NNs) for predicting human traits and diseases directly from large-scale genomics has shown that NN models perform worse than linear models. The findings show that the NNs were unable to capitalize on significant interaction effects. These are some examples: (a) Linear models can capture the upper bound of the genetic variance explained by genotyped SNPs, (b) For the traits tested, there are either not enough samples or not enough SNPs measured to find the interaction effects, (c) In the case of diseases, there is a significant risk of mislabeled samples, which could affect the model's ability to learn subtle interaction effects, (d) Interaction effects are present, but their effects are non-significant, and (e) Despite the fact that significant non-linear interactions have not been widely demonstrated, there are still advantages to developing NN-based models for disease risk prediction. One factor is the adaptability of NN-based models, which can be built to accept multi-modal inputs and predict multiple traits at the same time. Some of these inputs, such as images or text, may be unstructured and highly non-linear. This alone is a significant advantage over linear models. Furthermore, while the individual modalities can be mostly modeled with linear models, there is always the possibility of complex interaction effects, such as between clinical measurements and other biological data. We can discover such relationships by using NNs.

However, there are numerous difficulties in developing complex NN models that can be applied to human health data. The vast amount of biological data is a major challenge. Genomic data, for example, frequently contain millions of genetic variants genotyped for large sample sizes. Another barrier to fully leveraging health data is that it is frequently multi-modal. Supervised machine learning tasks are frequently trained to accept a single type of input, such as identifying the main object in a given data. Health data, on the other hand, can include multi-omics data such as genomics, transcriptomics, and proteomics, as well as targeted biochemical and clinical data and even ultra-high resolution imaging.

As a result, we created a DL framework that can integrate large-scale genomics data with other omics or clinical data. Among the framework's features is a new neural network model called Genome-Local-Net (GLN), which we created specifically for large-scale genomics data. GLN is based on a custom Locally-Connected Layer (LCL) that we created, and it extracted genetic information from genome-wide data with comparable or better

Correspondence to: Franco Ferrari, Department of Biotechnology, Illinois university, Chicago, USA, E-mail: fferrari@iit.edu

Received: 02-May-2022, Manuscript No. JDMGP-22-17225; **Editor assigned:** 04-May-2022, PreQC No. JDMGP-22-17225 (PQ); **Reviewed:** 18-May-2022, QC No JDMGP-22-17225; **Revised:** 25-May-2022, Manuscript No. JDMGP-22-17225 (R); **Published:** 02-Jun-2022. DOI: 10.4172/2153-0602.22.13.253.

Citation: Ferrari F (2022) Integrating Genomics Improves Large-Scale of Human Genomics. J Data Mining Genomics Proteomics. 13: 253.

performance than the other NN models we tested. We discovered that GLN performed statistically better overall on 338 diseases, disorders, and traits than our implementation of the Least Absolute Shrinkage and Selection Operator (LASSO).

Autoimmune diseases, such as Type 1 Diabetes (T1D) and rheumatoid arthritis, were of particular interest because they have previously been shown to have complex interaction effects. A detailed examination of the T1D SNPs most strongly activated by the GLN model reveals extensive interactions between them, even across chromosomes. All models in the framework automatically extend to Multi-Task (MT) learning, which we demonstrate by training a single GLN model topredict 338 diseases simultaneously. Furthermore, when training GLN-based models across 290 diseases in the UK Bio bank, we incorporate genotype data, age and gender covariates, blood measurements, urine measurements, and various anthropometrics. The integration of these measurements demonstrates a significant improvement in almost all traits, highlighting the potential of integrative models for health-based predictions. We use explainable AI to identify relevant SNPs and clinical measurements that are consistent with disease literature, and we show that combining genotype and clinical data results in better calibrated models for precision medicine.