Big Data in Biology: what the pandemic has taught us

Ewan Birney

European Molecular Biology Laboratory, European Bionformatics Institute, UK

Abstract: Molecular biology is now a leading example of a data intensive science, with both pragmatic and theoretical challenges being raised by data volumes and dimensionality of the data. These changes are present in both "large scale" consortia science and small scale science, and across now a broad range of applications – from human health, through to agriculture and ecosystems. All of molecular life science is feeling this effect. This shift in modality is creating a wealth of new opportunities and has some accompanying challenges. In particular there is a continued need for a robust information infrastructure for molecular biology. This ranges from the physical aspects of dealing with data volume through to the more statistically challenging aspects of interpreting it. A particular problem is finding causal relationships in the high level of correlative data. Genetic data are particular useful in resolving these issues. The pandemic has brought together operational public health delivery (eg, testing and DNA sequencing of the infectious agent) alongside research and models. The rate of learning has increased between these two domains and delivered better and better products for both policy makers and research. I will illustrate this with examples including the expansion of the Alpha and Delta SARS-CoV-2 genomes and integrating genomic and contact tracing work.