The bioinformatics challenge of 2 million genomes

Mark A. Collins[a], Marc Flesch[b], Michael Remmer[b], Daniel Nesbit[a] & Tamas Rujan[c]
[a]Genedata Inc, Lexington, USA | [b]Genedata GmbH, Munich, Germany | [c]Genedata AG, Basel, Switzerland

Recent announcements by pharmaceutical companies[1] and even governments[2] regarding large scale "moon-shot" genomics initiatives to improve human health got us thinking here at Genedata. So we sat our experts down to discuss what it would take to leverage the full power of two million genomes. They came up with seven key focus areas from the informatics perspective which we review below.

Efficiency and scalability of data storage, management and processing—2 million genomes equates to around 150 petabytes of genomic data alone (raw BAM files etc), plus additional storage for clinical and phenotype data, with a growth rate of 5Pb per year. If the trend towards sequencing as many people as possible continues, some estimates point to 2 billion genomes within 10 years, which would be multiple exabytes of data[3]. Storage will have to be hierarchical with different types of physical storage, compression levels, and latency, as well as level of detail of the data. In turn, highly efficient data management—probably involving some form of hybrid architecture—that leverages elasticity of cloud computing with the performance and lower costs of on-premise, high performance computing clusters will be needed.

Furthermore, it will be impractical to move data; computer processing power will have to be brought to the data. Therefore, data needs to be linked and federated from the source and new ways to process data that use bandwidth efficiently will be required.

Harmonizing disparate data—deriving scientific insights from genomic and other omic data requires further data, such as phenotype data, medical records, treatment histories, data in public databases, clinical trials etc. All this data will need to be brought together for analyses, hence technologies that allow federation and semantically correct curation of data using strong dynamic ontologies will be key here. In harmonizing data it will also be key to understand where data has come from (data provenance), the chain of custody and what has been done to the data (audit trail).

Fast query performance—gaining value from this kind of data will demand that queries, both of phenotypes and genotypes, can be executed quickly and efficiently. The best way to do this is up for debate, e.g. graph databases, columnar databases, noSQL etc. Furthermore, query tools will need to understand the context of the data, so ontologies and semantic approaches may well be needed, as will hierarchical and distributed query technologies.

Extensibility—technologies change, so it is important that from both an omic and associated clinical data perspective the infrastructure be able to accommodate new data types and changes to existing data. For example, can the data model accommodate array data and NGS data? What about other omics such as mass spec data? In addition, assessment of the impact of long read technologies on data acquisition, storage and processing should be considered. Furthermore, as data is from clinical studies, time-series, longitudinal and numeric, categorical data will need to be handled.

Security and Privacy—the data being collected is from patients and while it has huge benefit for the development of new therapies, ensuring that the data from those individuals who have consented to its collection is kept safe is a key trust issue. Strong encryption, the ability to anonymize data and

ensuring that the person donating data has control over its use, both now and in the future, will be key considerations. Since 2 million genomes will likely be from global populations, the regulatory and legal considerations of moving data across international borders will have to be addressed, as will ethical obligations to return research results to individuals.

Collaboration—a successful outcome from 2 million or more genomes requires input from potentially thousands of different individual stakeholders, from bench scientists, through bioinformaticians, to computational biologists and translational researchers. Any infrastructure has to be able to ensure that each stakeholder can work on the data in an efficient manner, at scale and in compliance with an ever changing regulatory climate.

Scientific Insights—the scale, breadth and depth of the data represent a priceless opportunity. Ensuring the maximum scientific benefit from this is really where the "rubber meets the road". Any infrastructure will have to be able to process, analyze and visualize data, leveraging the latest developments in information theory, statistics, computational biology and machine learning (e.g. deep learning). Furthermore, scientific insights demand an understanding of data quality so new ways to rapidly determine quality metrics prior to analysis is also a core requirement. As data is analyzed for different outcomes and by different stakeholders these quality metrics must accompany the data to ensure overall robustness of any decisions made.

Summary

There is no doubt that genomics data is the poster child for big data and will likely exceed other sources (e.g. social media, astronomy and video), in terms of the challenge of data acquisition, storage, distribution and analysis[3]. To maximize the benefit to human health of 2 million genomes or more will require a huge global interdisciplinary effort with contribution from pharmaceutical companies, academic institutes, and instrument, reagent and software providers. At Genedata we have spent the last 20 years working together with leading life science organizations to help them gain scientific insights from complex, data rich R&D processes. One of the ways we do this is by providing Genedata Profiler™, an enterprise software solution that empowers researchers to generate valuable scientific insights from omic profiling of patients through "industrializing" the processing, analysis and management of huge volumes of omic data, such as that from 2 million genome type projects.

References

1. AstraZeneca launches integrated genomics approach to transform drug discovery and development, April 2016
2. Edith Mitchell. (2016) Moonshot Toward a Cure for Cancer. Journal of the National Medical Association 108:2, 104-105
3. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.100

To learn more visit us on the web at www.genedata.com/profiler/ or contact us at profiler@genedata.com.