European Journal of Human Genetics (2017) 25, 656–657

www.nature.com/ejhg

BOOK REVIEWS

How to manage large-scale collaborative genomics research projects?

Collaborative Genomics Projects: A Comprehensive Guide
Edited by: Margi Sheth, Jiashan Zhang and Jean C Zenklusen
ISBN: 9780128021439
Published by: Elsevier, 2016
Price: £54.99/€64.95/$89.95

European Journal of Human Genetics (2017) **25,** 656; doi:10.1038/ejhg.2017.19

Collaborative Genomics Projects: A Comprehensive Guide establishes a framework for managing large-scale genomic research projects involving multiple collaborators, describes lessons learned through The Cancer Genome Atlas (TCGA) project, and suggests efficient project management strategies.

All authors of this book are contributors to TCGA Program, an NIH-funded effort to generate a comprehensive catalogue of genomic alterations in more than 35 types of cancer. This international project united seven institutions and was primarily carried out in the USA. The lessons learned in this project are widely applicable and especially useful for teams looking to set up or collaborate within a large-scale genomics research project.

As the cost of genomic sequencing decreases, more and more researchers are using genomic data to explain various aspects of disease and health biology. The amount of genomic data generated is growing exponentially. Protocols are needed for the long-term storage, dissemination and regulation of these data for research. This book can be regarded as a complete handbook on the management of research projects involving genomic data.

Chapters from 4 to 7 analyse the establishment of data generation and analysis pipelines, data storage and dissemination, quality control, auditing, and reporting. The reader is walked through the technological challenges faced by data centres and hubs that ingest and distribute hundreds of terabytes monthly to over 8000 unique end users. The data have to be controlled for quality, properly labelled, protected, and analysed. Dissemination of the research findings generated by large scientific consortia can be challenging, and it requires careful planning of the intended audience, scientific journals, and authorship and embargo policies. Each chapter is concluded with references, including appropriate regulatory standards and policy documents.

In chapter 5, the authors discuss a data generation pipeline that should be designed very early in the process. In building the data generation pipeline, three important aspects have to be determined: (1) which types of genomic data should be generated to accomplish the goals of the project: for example, in the 1000 Genomes project the goal was to create a catalogue of structural variants across diverse populations, and therefore whole-genome sequencing was performed; (2) what technologies should be used to generate data (exome/genome sequencing, ChIP-seq, RNA-Seq, etc.): the level of processing required to present the data, whether raw, processed or interpreted; (3) who will be generating data: centralised or federated, single or multi-centre.

Chapter 6 provides an in-depth analysis of requirements for formats of data exchange, storage and dissemination. A large-scale genomic project will create huge amounts of data that have to be centrally managed. This function is performed by a data management or coordination centre (DCC). The DCC defines data management work flows and standardisation, and formatting of data and meta-data, and performs quality control of incoming data. In addition, experimental protocols and analytical tools used to derive the data should be available in a standardised format that ensures reproducibility. The DCC makes sure that its data management system complies with all federal, state, local and institutional data security policies. The data management principles covered by this chapter are widely applicable, even though they were described as pertaining to TCGA.

Chapter 7 focuses on data analysis. A starting point in data analysis is a list of perceived scientific questions that the data might answer. These questions determine the most effective analytical approach. In TCGA it was clustering. Data analysis was performed by multi-disciplinary teams of about 1500 individuals. Working groups tackled specific scientific questions that were later published. The coordination of publishing with a large number of collaborators was challenging. The chapter concludes with an in-depth analysis of appropriate strategies of publishing in high-impact scientific journals that principal investigators and team leaders should be aware of.

This timely book will be very interesting for large-scale genomic project managers and individual researchers. It is essential for those who plan research infrastructure, as well as team leaders of large consortia and research policy makers.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Erinija Pranckevičienė and Vaidutis Kučinskas
Human Genome Research Centre, Vilnius University, Vilnius, Lithuania
V Kučinskas,
E-mail: vaidutis.kucinskas@mf.vu.lt