
Preface

Next-generation DNA sequencing (NGS) is perhaps the most dynamic and rapidly growing area of modern biology. To keep pace with the rapid developments in sequencing technology, innovative biological applications, and software, it has been necessary to extensively revise the first edition of this book published in 2013. As the cost of sequencing continues to fall and the scope of sequencing projects expands to larger studies and advances into new areas (ecology, taxonomy, population genetics, and chromatin structure, to name a few), data processing and analysis remain key challenges. In addition to extensive revisions capturing new publications in every chapter, new chapters have been added on emerging NGS technologies, quality control, de novo transcriptome assembly, cloud-based informatics, and the nascent field of proteogenomics.

NGS is enabled by sophisticated and novel bioinformatics tools specifically created or adapted to make NGS possible. Not only has new software been developed for a wide range of novel applications and types of data analysis, but new algorithms have also been developed for old problems, such as sequence alignment and de novo assembly, to cope with the huge volumes of data generated on new sequencing machines. The cycle of software development is very rapid as vendors upgrade their machines and different groups compete to publish new methods and to meet investigator demands. As a result of the frenetic pace of development, new software tools for NGS data analysis are often released with bare bones command line user interfaces and minimal documentation. Making things even more complicated, many different software packages exist for each of the major NGS applications with few benchmarking studies available to guide users in the choice of the best solutions. In short, there is an urgent need for a practical guide for researchers about all major aspects of informatics needed to successfully operate and fully take advantage of NGS.

Many of the authors contributing to this book have been based at the New York University Langone Medical Center (NYULMC), which has invested early and heavily in building both assay and informatics capacity and manpower in NGS. Specifically, NYULMC built in 2008 its Genome Technology Center to provide basic and

translational scientists access to the latest DNA sequencing, expanding upon previous technologies such as microarrays and real-time polymerase chain reaction (qPCR). In parallel, an NGS Informatics Group was created to provide research design, upstream data processing, data management, and data analysis consulting for all users of the sequencers within NYULMC and beyond. As the Informatics Group has grown in experience and expanded to include many collaborators, it has tested many different software packages and built best-practice workflows for many NGS projects, including *de novo* sequencing (and genome annotation), amplicon and exon sequencing for variant detection and for metagenomics, ChIP-seq, RNA-seq, and detection of somatic variants in cancer.

In this book, building on our own extensive experience that spans collaborations on more than 30 National Institutes of Health-funded projects, and by critically evaluating and synthesizing the literature in the field, we provide an overview of many core types of NGS projects, a discussion of methods embodied in popular software, and detailed descriptions of our own best-practice workflows (including several tutorials). We have included advice designed to be helpful to both bioinformaticians implementing their own data analysis methods and to laboratory and clinical investigators planning to use NGS methods to address their own research questions. The future of NGS and all the related informatics innovations is as bright as it is exciting, and we are gratified to be able to contribute to the field's development with the present updated volume.

STUART M. BROWN