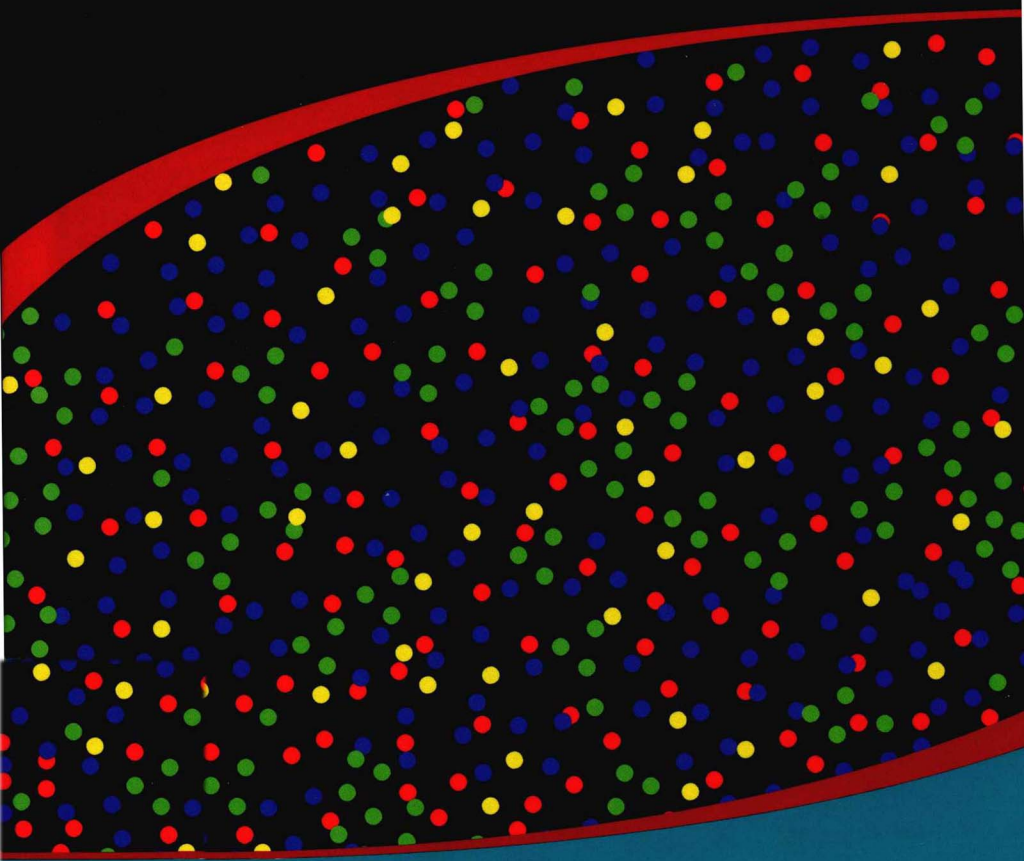
Bioinformatics
A Practical Handbook of Next Generation Sequencing and Its Applications

Lloyd Low • *Martti Tammi* editors



World Scientific

Bioinformatics
A Practical Handbook of Next Generation Sequencing and Its Applications

editors

Lloyd Low
Perdana University Centre for Bioinformatics, Malaysia
The Davies Research Centre, University of Adelaide, Australia

Martti Tammi
Sime Darby, Malaysia

World Scientific

NEW JERSEY · LONDON · SINGAPORE · BEIJING · SHANGHAI · HONG KONG · TAIPEI · CHENNAI · TOKYO

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Names: Low, Lloyd, editor. | Tammi, Martti (Martti T.), editor.

Title: Bioinformatics : a practical handbook of next generation sequencing and its applications / edited by Lloyd Low (Perdana University Centre for Bioinformatics, Malaysia) and Martti Tammi (Sime Darby, Malaysia).

Description: New Jersey : World Scientific, 2016. | Includes bibliographical references and index.

Identifiers: LCCN 2016040510 | ISBN 9789813144743 (hardcover : alk. paper)

Subjects: LCSH: Bioinformatics. | Nucleotide sequence. | Molecular biology. | Gene mapping.

Classification: LCC QH324.2 .B547125 2016 | DDC 570.285--dc23

LC record available at https://lccn.loc.gov/2016040510

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Typeset by Stallion Press

Email: enquiries@stallionpress.com

Printed in Singapore

Bioinformatics
A Practical Handbook of Next Generation Sequencing and Its Applications

Foreword

Olivo Miotto

Several years ago, when the first draft of the human genome was being completed, I decided to focus my efforts on the study of pathogen genomes. Armed with a background in software engineering, one of the first things that preoccupied me was a problem that loomed on the horizon and had little to do with the fascinating biology that was emerging from the study of genomes. It was already clear that, in order to study genetic variations, their effects on phenotype, and their epidemiological dynamics, it would be necessary to collect massive amounts of data, far more than most of us could actually handle. The question was not so much whether storage or processing capabilities would be sufficient — Moore's Law had accustomed us to rapid growth in computing power, and I was confident these technical challenges could be met. The critical question was whether the people who would be analysing these data would have sufficient know-how and resources to handle these large quantities of data, and extract the knowledge they needed. To be sure, the same problem was faced by companies that needed to build search engines, hotel booking systems, web-based ratings software, and all the services based on what we now call "big data". But genomics looked like a problem that could not be tackled by computer scientists alone. Biologists had to be empowered to handle scary amounts of data.

Those issues were evident even before whole-genome sequencing was revolutionized by the next-generation sequencing (NGS) technologies introduced by companies such as Solexa (now Illumina). Today, the MalariaGEN genomic epidemiology project on which I work (malariagen.net/projects/p-falciparum-community-project) comprises the genomes of

Plasmodium parasites from almost ten thousand clinical samples, each backed by several gigabytes of short-read sequencing data — far more data than I would have predicted a few years ago. And yet, the knowledge gap has not been properly filled: if anything, it has become increasingly harder for life scientists and clinicians to effectively process such massive quantities of data, and many projects rely on collaborations with informatics specialists who often have limited expertise of the biological domain.

In the light of these difficulties, I give full credit to Lloyd Low and Martti Tammi for making a significant contribution towards filling the gap. What they have produced is a very practical guide, part reference and part tutorial, that will be appreciated by many life scientists for its direct and straightforward approach. Crucially, the content of this book is based on years of teaching experience, and "fine-tuned" by keeping in mind the difficulties routinely faced by those learning how to deal with NGS data. It contains a useful toolkit of techniques and practices using some of the most popular tools in use, such as BWA, samtools and so on.

The material covered in this book will support a broad range of applications: the final chapter suggests some possibilities, but clearly each reader will have to tackle challenges unique to their own areas of study, and this work will serve as a base on which to build further techniques. Commendably, it promotes the definition of a well-organized analytical workflow, and gives prominence to the quality aspects of genomics work — hugely important and frequently underestimated. Conducting a GWAS — or constructing a phylogeny — without first properly evaluating what data to rely upon and what to discard will invariably lead to useless or false results. It is therefore essential to instil high standards of quality into the mind of students and anyone undertaking genomic analyses.

I wish all readers all the best in their endeavours in this complex field, which I hope they will find rich in rewards.

Olivo Miotto

Mahidol-Oxford Tropical Medicine Research Unit, Bangkok, Thailand
Wellcome Trust Sanger Institute, Hinxton, United Kingdom
Centre for Genomics and Global Health, Oxford University, United Kingdom

Foreword

Nazar Zaki

The revolution that Next Generation Sequencing (NGS) brought to genetics can be compared to the revolution that the invention of the telescope brought to astronomy. Genetic phenomena can now be studied at the molecular level and genetic processes can be studied at genomic, transcriptomic and epigenomic levels using NGS technologies. The low cost of sequencing is allowing human genomes to be sequenced routinely and longer sequencing lengths allow the easier construction of novel genomes. Therefore, it is essential for researchers working in biology to have a good grasp of basic concepts in handling NGS data at different levels. This book provides a succinct and easy to read introduction to the processing of NGS data at various levels for a general audience.

For the novice user, the first three chapters provide a brief primer to the technology behind NGS and how to get past the hurdle of aligning NGS data to a reference genome. The alignment step is demonstrated using the popular open source aligner BWA and the commercial NovoAlign aligner that is known for its high accuracy. This chapter is written by an engineer at Novocraft itself and the reader can customize the workflow to achieve the required degree of precision and speed using NovoCraft products or open source options.

Once past the hurdle of aligning the reads, this book answers what naturally comes into mind: "What do I do next"? It introduces IGV so that the users can visualize the alignments and as the next step introduces the Galaxy framework to create a research workflow. Even if the user is not an expert in computer science, Galaxy will empower him to establish some basic research tasks after some experimenting. Overall, the reader

can start diving deeper into analysing NGS data on his own after reading the first five chapters of the book.

While most of the NGS analysis currently starts with alignment, there are other applications that require genome assembly. This is especially true for smaller genomes and it is becoming popular as NGS technologies that produce very long read lengths are made available. In future it may be the case that the borderline between sequence alignment and assembly will not be clear cut. In Chapter 6, Dr. Tammi shares his expertise on sequence assembly with a gentle introduction to the basics of sequence assembly. Not only does he show the reader how to assemble a genome, but he also teaches how to gauge the quality of an assembly.

In the next few chapters, the book concentrates on specific application of NGS. The book has picked a timely set of applications that are being widely used and the user is guided step-by-step on how to process data for each application. Exome sequencing has become an important branch of NGS due to its cost considerations and the higher depth of coverage. We also have the ability to take snapshots of cells in action using transcriptome sequencing. Another different branch that is benefitted by NGS is metagenomics, which tries to find answers about the total genomic content of samples in contrast to the previous applications we discussed. Another important question is how to extract the relationships between genotypes and phenotypes. All these applications need different approaches and asks different types of questions. However, techniques used in these areas can be carried over to other methods. For example, techniques used for processing exome sequencing can be useful in working with other targeted sequencing methods and techniques used to find variations in WGS can be used in transcriptomic studies. Therefore, the reader can benefit by understanding the concepts used to process these different types of data sets.

Nazar Zaki

Professor
Leader, Bioinformatics Research Team
Coordinator, Intelligent Systems
Coordinator, Software Development
College of Information Technology
United Arab Emirates University
Al Ain 17551, UAE

Preface

The secret of life is encoded in DNA sequences. Since the 1970s, many inventors and innovators have enhanced DNA sequencing technologies to enable us to move from the painstaking process of reading a single base to now being able to easily gather the sequences of millions of DNA fragments. Today, we live in the era where next generation sequencing (NGS) technologies are commonly available and third generation sequencers have also been commercialized. New users of NGS usually have not worked with Sanger sequenced data and their introduction to this field is a straight jump into a dizzying amount of sequences. It is an understatement to say that it is difficult to handle the massive amount of sequenced data and to use them to make biological discoveries.

The idea for this book was conceived after my colleagues and I had organized and taught at various workshops on NGS. We thought that it would be a great idea to provide a comprehensive practical oriented book on NGS so that more people can learn how to handle bioinformatics data that are coming from this technology. The book covers general topics on how to handle NGS data from sequence quality inspection, alignment of reads to finding single nucleotide polymorphisms (SNPs). Other advanced topics such as genome assembly, exome sequencing, transcriptomics, and metagenomics are also covered. A special last chapter is dedicated to applications of NGS data to give readers a taste of the power of this technology in genetic mapping and genome wide association studies (GWAS).

There are common difficulties faced by many first time learners who need to analyze NGS data. This book put together materials and experiences gained from teaching many first time learners and it includes

additional resources aimed at strengthening the readers knowledge in this field. We anticipate that this book will be of great use to students and researchers in the life sciences. For readers who are already proficient in NGS based data analysis, they can still keep the book as a reference material.

Note to readers: Companion datasets can be downloaded at http://bioinfo.perdanauniversity.edu.my/infohub/display/NPB/Index

Acknowledgements

First and foremost, I must thank Dr. Asif Khan of the Perdana University School of Data Science (PUSDS) for encouraging me to pursue writing a book on NGS. In addition, I am thankful for the continuous and steady support given by other staff and students at PUSDS. Two of them, Dr. Adeel Malik and Muhammad Farhan are also authors of the book. I also wish to thank Dr. Sean Mayes and Dr. David Ross Appleton for their reviews on various chapters. Last but not least, I wish to thank authors from Sime Darby Technology Centre, Novocraft and Institute of Statistics (Jakarta) for contributing book chapters. Without these key people, the book would not have been possible.

Lloyd Low

Contents

Chapter 1

Introduction to Next Generation Sequencing Technologies

Lloyd Low[a] and Martti T. Tammi[b]

[a]Perdana University Centre for Bioinformatics (PU-CBi),
Block B and D1, MAEPS Building, MARDI Complex,
Jalan MAEPS Perdana, 43400 Serdang, Selangor, Malaysia.
[b]Biotechnology & Breeding Department,
Sime Darby Plantation R&D Centre, Selangor, 43400, Malaysia.

A Brief History of DNA Sequencing

In 1962 James Watson, Francis Crick and Maurice Wilkins jointly received the Nobel Prize in Physiology/Medicine for their discoveries of the structure of deoxyribonucleic acid (DNA) and its significance for information transfer in living material.[1] The secret of DNA in orchestrating living activities lies in the arrangement of the four bases (i.e. adenine, thymine, guanine and cytosine). The linear sequence of the four bases can be considered as the language of life with each word specified by a codon that is made up of three bases. It was an interesting puzzle to figure out how codons specify amino acids. In 1968, Robert W. Holley, HarGobind Khorana and Marshall W. Nirenberg were awarded the Nobel Prize in Physiology/Medicine for solving the genetic code puzzle. Now it is known that collection of codons direct what, where, when and how much proteins should be made. Since the discovery of the structure of DNA and the genetic code, deciphering the meaning of DNA sequences has been an ongoing quest by many scientists to understand the intricacies of life.

The ability to read a DNA sequence is a prerequisite to decipher its meaning. Not surprisingly then, there has been intense

competition to develop better tools to sequence DNA. In the 1970s, the first revolution in DNA sequencing technology began and there were two major competitors in this area. One was the commonly known Sanger sequencing method[2,3] and another was the Maxam–Gilbert sequencing method.[4] Over time, the popularity of the Sanger sequencing method and its modifications grew so much that it overshadowed other methods until perhaps 2005 when Next Generation Sequencing (NGS) began to take off.

In 1977, Sanger and colleagues successfully used their sequencing method to sequence the first DNA-based genome, a ϕX174 bacteriophage, which is approximately 5375 bp.[5] This discovery heralded the start of the genomics era.Initially, the Sanger sequencing method in 1975 used a two-phase DNA synthesis reaction.[2] In the first phase, a DNA polymerase was used to partially extend a primer bound onto a single stranded DNA template to generate DNA fragments of random lengths. In phase two, the partially extended templates from the earlier reaction were split into four parallel DNA synthesis reactions where each reaction only had three of the four deoxyribonucleotide triphosphates (dNTPs; which is made up of dATP, dCTP, dGTP, dTTP). Due to a missing deoxyribonucleotide triphosphate (e.g. dATP), the DNA synthesis reaction would stop at its 3′ end position just one position prior to where the missing base was supposed to be incorporated. All of these synthetized DNA fragments could then be separated by size using electrophoresis on an acrylamide gel. The DNA sequence could be read off a radioautograph since its DNA synthesis happened with the incorporation of radiolabeled nucleotides (e.g. S-dATP).[35]

There were many problems with the initial version of the Sanger sequencing method that required further innovations before its widespread use and this scenario is akin to what is happening in the recent NGS technological developments. Some problems of the early Sanger sequencing method included the cumbersome two-phase procedures, only short length of a DNA sequence could be determined, the requirement of a primer meant some sequences of the template had to be known, hazardous radio labeled nucleotides were used and there was also no automated

way to read off a DNA sequence. Sanger and colleagues rapidly improved on the method described in 1975 by eliminating the two-phase procedure with the use of dideoxynucleotides as chain terminators.[3] Briefly, the improved method started with four reaction mixtures that already had the single stranded DNA template hybridized to a primer. In each reaction, the DNA synthesis proceeded with four deoxyribonucleotide triphosphates (one with radiolabeled nucleotide) and one dideoxynucleotide (ddNTP). Whenever a dideoxyribonucleotide was incorporated, the reaction terminated and thereby produced a mixture of truncated fragments of varying lengths. These DNA fragments were then separated by electrophoresis and then read off from aradioautograph. By adjusting the concentration of ddNTPs, chain termination can be manipulated to produce a longer sequence read.

To solve the requirement of knowing some template sequences for primer design, cloning was introduced. For example, the M13 sequencing vector is commonly used as a holder for DNA insert and known primers that bind to the vector sequence are available to be used to sequence the unknown DNA insert. One major innovation to the Sanger sequencing method is the replacement of radioactive labels with fluorescent dyes.[6] Four different dye colour labels are available for the four dideoxynucleotide chain terminators and thus, DNA fragments that terminate at all four bases can be generated in a single reaction and thus analyzed on a single lane of acrylamide gel. The electrophoresis is coupled to a fluorescent detector that is also connected to a computer and thus sequence data can be automatically collected. In 1986, Applied Biosystems commercialized the first automated DNA sequencer (i.e. Model 370A) that is based on the Sanger sequencing method. For an animation of the Sanger sequencing method, the reader should refer to the Welcome Trust Sanger Institute (http://www.wellcome.ac.uk/Education-resources/Education-and-learning/Resources/Animation/WTDV026689.htm).

Due to limitations of the chain terminator chemistry and resolution of the electrophoresis method, the Sanger sequencing method is only capable of sequencing a read of about 500 to 800 bases long. Most genes and other interesting DNA sequences are

longer than that. Therefore, a method is required to first break up a longer DNA molecule into fragments, sequence the individual fragments and then piece them together to create a contiguous sequence (i.e. contig). In one approach known as the shotgun sequencing, the long DNA fragment is randomly sheared and then cloned for sequencing.[7] A computer program is then used to assemble the sequences by finding overlaps. It is challenging to find sequence overlaps when thousands to millions of DNA fragments are generated. The problem requires alignment algorithms and some notable examples of early work in this area include the Needleman-Wunsch algorithm[8] and Smith-Waterman algorithm.[9] Details on the bioinformatics involved in NGS alignment tools and sequence assembly are given in Chapters 4 and 6, respectively.

Next Generation Sequencing Technologies

One of the goals of the Human Genome Project (HGP) is to support advancements in DNA sequencing technology.[10] Although the HGP was completed with the Sanger sequencing method, many groups of researchers were already tinkering with new ideas to increase throughput and decrease cost of sequencing prior to the announcement of the first human genome draft in 2001. For example, developments for nanopore sequencing can be traced back to 1996 when researchers experimented with α-hemolysin.[11] After years of experimentations, the second DNA sequencing technology revolution finally took off in 2005 and ended Sanger sequencing dominance in the marketplace. The revolution is still ongoing at the time of this writing and it can be seen from the rapid decline in the cost of sequencing since the introduction of NGS technologies (Figure 1).

The sequencing technologies associated with the second revolution are referred to by various names, including second generation sequencing, NGS and high throughput sequencing. It should perhaps be most appropriately termed as high throughput sequencing but NGS seems to be more commonly used to categorize such technologies and hence, this term is used for the book. For the purpose of this book, NGS technology refers to platforms that are