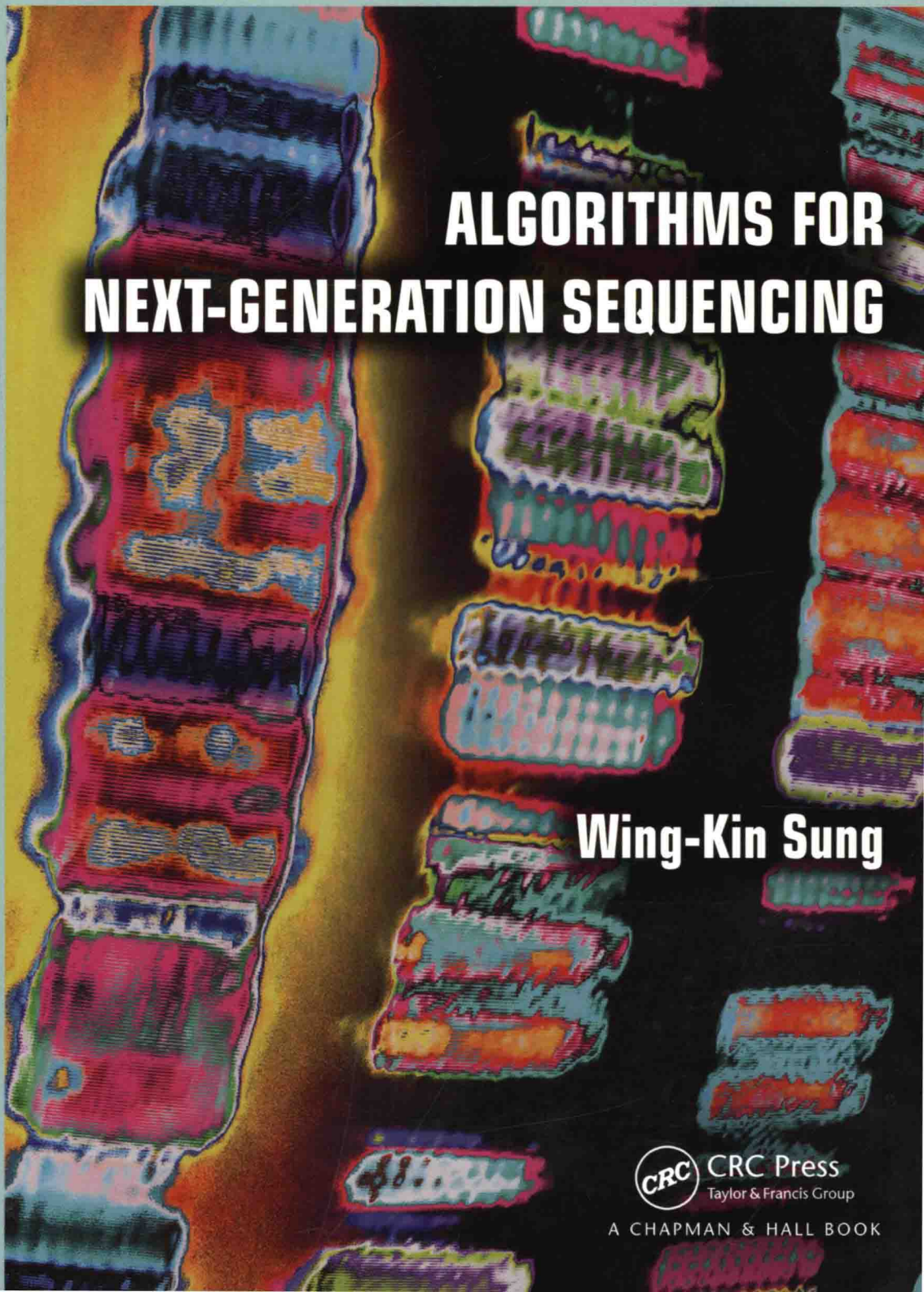
Chapman & Hall/CRC
Mathematical and Computational Biology Series



# ALGORITHMS FOR NEXT-GENERATION SEQUENCING

Wing-Kin Sung

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Mathematical and Computational Biology Series

# ALGORITHMS FOR NEXT-GENERATION SEQUENCING

Advances in sequencing technology have allowed scientists to study the human genome in greater depth and on a larger scale than ever before – as many as hundreds of millions of short reads in the course of a few days. But what are the best ways to deal with this flood of data?

**Algorithms for Next-Generation Sequencing** is an invaluable tool for students and researchers in bioinformatics and computational biology, biologists seeking to process and manage the data generated by next-generation sequencing, and as a textbook or a self-study resource. In addition to offering an in-depth description of the algorithms for processing sequencing data, it also presents useful examples illustrating how the algorithms work.

**Features**

- One of the first books published on this key topic
- Written by a leading practitioner
- Focuses on algorithms
- Covers technologies used in next-generation sequencing
- Includes a wide range of case studies and applications

**Wing-Kin Sung** is a professor in the Department of Computer Science of the National University of Singapore and a senior group leader in the Genome Institute of Singapore. He has over 20 years of experience in algorithm and bioinformatics research.

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K16105
ISBN: 978-1-4665-6550-0
90000
9 781466 565500
www.crcpress.com

# ALGORITHMS FOR NEXT-GENERATION SEQUENCING

Sung

Chapman & Hall/CRC Mathematical and Computational Biology Series

# ALGORITHMS FOR NEXT-GENERATION SEQUENCING

Wing-Kin Sung


CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742



Printed at CPI on sustainably sourced paper
Version Date: 20170421

International Standard Book Number-13: 978-1-4665-6550-0 (Hardback)



**Visit the Taylor & Francis Web site at**
**http://www.taylorandfrancis.com**

**and the CRC Press Web site at**
**http://www.crcpress.com**

# ALGORITHMS FOR NEXT-GENERATION SEQUENCING

# *Preface*

Next-generation sequencing (NGS) is a recently developed technology enabling us to generate hundreds of billions of DNA bases from the samples. We can use NGS to reconstruct the genome, understand genomic variations, recover transcriptomes, and identify the transcription factor binding sites or the epigenetic marks.

The NGS technology radically changes how we collect genomic data from the samples. Instead of studying a particular gene or a particular genomic region, NGS technologies enable us to perform genome-wide study unbiasedly. Although more raw data can be obtained from sequencing machines, we face computational challenges in analyzing such a big dataset. Hence, it is important to develop efficient and accurate computational methods to analyze or process such datasets. This book is intended to give an in-depth introduction to such algorithmic techniques.

The primary audiences of this book include advanced undergraduate students and graduate students who are from mathematics or computer science departments. We assume that readers have some training in college-level biology, statistics, discrete mathematics and algorithms.

This book was developed partly from the teaching material for the course on Combinatorial Methods in Bioinformatics, which I taught at the National University of Singapore, Singapore. The chapters in this book are classified based on the application domains of the NGS technologies. In each chapter, a brief introduction to the technology is first given. Then, different methods or algorithms for analyzing such NGS datasets are described. To illustrate each algorithm, detailed examples are given. At the end of each chapter, exercises are given to help readers understand the topics.

Chapter 1 introduces the next-generation sequencing technologies. We cover the three generations of sequencing, starting from Sanger sequencing (first generation). Then, we cover second-generation sequencing, which includes Illumina Solexa sequencing. Finally, we describe the third-generation sequencing technologies which include PacBio sequencing and nanopore sequencing.

Chapter 2 introduces a few NGS file formats, which facilitate downstream analysis and data transfer. They include fasta, fastq, SAM, BAM, BED, VCF and WIG formats. Fasta and fastq are file formats for describing the raw sequencing reads generated by the sequencers. SAM and BAM are file formats

for describing the alignments of the NGS reads on the reference genome. BED, VCF and WIG formats are annotation formats.

To develop methods for processing NGS data, we need efficient algorithms and data structures. Chapter 3 is devoted to briefly describing these techniques.

Chapter 4 studies read mappers. Read mappers align the NGS reads on the reference genome. The input is a set of raw reads in fasta or fastq files. The read mapper will align each raw read on the reference genome, that is, identify the region in the reference genome which is highly similar to the read. Then, the read mapper will output all these alignments in a SAM or BAM file. This is the basic step for many NGS applications. (It is the first step for the methods in Chapters 6–9.)

Chapter 5 studies the de novo assembly problem. Given a set of raw reads extracted from whole genome sequencing of some sample genome, de novo assembly aims to stitch the raw reads together to reconstruct the genome. It enables us to reconstruct novel genomes like plants and bacteria. De novo assembly involves a few steps: error correction, contig assembly (de Bruijn graph approach or base-by-base extension approach), scaffolding and gap filling. This chapter describes techniques developed for these steps.

Chapter 6 discusses the problem of identifying single nucleotide variations (SNVs) and small insertions/deletions (indels) in an individual genome. The genome of every individual is highly similar to the reference human genome. However, each genome is still different from the reference genome. On average, there is 1 single nucleotide variation in every 3000 bases and 1 small indel in every 1000 bases. To discover these variations, we can first perform whole genome sequencing or exome sequencing of the individual genome to obtain a set of raw reads. After aligning the raw reads on the reference genome, we use SNV callers and indel callers to call SNVs and small indels. This chapter is devoted to discussing techniques used in SNV callers and indel callers.

Apart from SNVs and small indels, copy number variations (CNVs) and structural variations (SVs) are the other types of variations that appear in our genome. CNVs and SVs are not as frequent as SNVs and indels. Moreover, they are more prone to change the phenotype. Hence, it is important to understand them. Chapter 7 is devoted to studying techniques used in CNV callers and SV callers.

All above technologies are related to genome sequencing. We can also sequence RNA. This technology is known as RNA-seq. Chapter 8 studies methods for analyzing RNA-seq. By applying computational methods on RNA-seq, we can recover the transcriptome. More precisely, RNA-seq enables us to identify exons and split junctions. Then, we can predict the isoforms of the genes. We can also determine the expression of each transcript and each gene.

By combining Chromatin immunoprecipitation and next-generation sequencing, we can sequence genome regions that are bound by some transcription factors or with epigenetic marks. Such technology is known as ChIP-seq. The computational methods that identify those binding sites are known

as ChIP-seq peak callers. Chapter 9 is devoted to discussing computational methods for such purpose.

As stated earlier, NGS data is huge; and the NGS data files are usually big. It is difficult to store and transfer NGS files. One solution is to compress the NGS data files. Nowadays, a number of compression methods have been developed and some of the compression formats are used frequently in the literatures like BAM, bigBed and bigWig. Chapter 10 aims to describe these compression techniques. We also describe techniques that enable us to randomly access the compressed NGS data files.

Supplementary material can be found at

`http://www.comp.nus.edu.sg/~ksung/algo_in_ngs/`.

I would like to thank my PhD supervisors Tak-Wah Lam and Hing-Fung Ting and my collaborators Francis Y. L. Chin, Kwok Pui Choi, Edwin Cheung, Axel Hillmer, Wing Kai Hon, Jansson Jesper, Ming-Yang Kao, Caroline Lee, Nikki Lee, Hon Wai Leong, Alexander Lezhava, John Luk, See-Kiong Ng, Franco P. Preparata, Yijun Ruan, Kunihiko Sadakane, Chialin Wei, Limsoon Wong, Siu-Ming Yiu, and Louxin Zhang. My knowledge of NGS and bioinformatics was enriched through numerous discussions with them. I would like to thank Ramesh Rajaby, Kunihiko Sadakane, Chandana Tennakoon, Hugo Willy, and Han Xu for helping to proofread some of the chapters. I would also like to thank my parents Kang Fai Sung and Siu King Wong, my three brothers Wing Hong Sung, Wing Keung Sung, and Wing Fu Sung, my wife Lily Or, and my three kids Kelly, Kathleen and Kayden for their support.

Finally, if you have any suggestions for improvement or if you identify any errors in the book, please send an email to me at `ksung@comp.nus.edu.sg`. I thank you in advance for your helpful comments in improving the book.

Wing-Kin Sung

# *Contents*