

METHODS IN MOLECULAR BIOLOGY

Series Editor

John M. Walker

School of Life and Medical Sciences

University of Hertfordshire

Hatfield, Hertfordshire AL10 9AB, UK

For further volumes:

<http://www.springer.com/series/7651>

Insect Genomics

Methods and Protocols

Edited by

Susan J. Brown

Division of Biology, Kansas State University, Manhattan, NY, USA

Michael E. Pfrender

Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA

 **Humana Press**

Editors

Susan J. Brown
Division of Biology
Kansas State University
Manhattan, NY, USA

Michael E. Pfrender
Department of Biological Sciences
University of Notre Dame
Notre Dame, IN, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-8774-0 ISBN 978-1-4939-8775-7 (eBook)
<https://doi.org/10.1007/978-1-4939-8775-7>

Library of Congress Control Number: 2018958311

© Springer Science+Business Media, LLC, part of Springer Nature 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Humana Press imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.

The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

Insects surround us and play important roles in our ecosystem, from pollinating fruits to vectoring disease. As the founders of the initiative to sequence five thousand genomes (i5K) emphasized, understanding insect genomes will inform insect biology at all levels, from molecular to physiological. When scientists focused on physiology, toxicology, or development decide to sequence the genome of their favorite organism, they must address a raft of questions before they can start, because every step in a genome project presents unique challenges. The content of this volume focuses on methods currently used in sequencing, assembling, and analyzing insect genomes to help graduate students, postdocs, and PIs that are new to genomics navigate this new terrain. (The target audience includes graduate students, postdocs, and research scientists new to the field.)

There are many different ways to approach a genome project, and several strategies are presented in Chapter 1. Genome size must be determined to guesstimate the amount of data required to produce a complete assembly. A protocol for determining genome size by fluorescent measurement in a flow sorter is described in Chapter 2. Although there are specific DNA isolation kits described for most next-generation sequencing technologies that produce ~50 kb DNA, there are circumstances that required much longer starting material, and a HMW DNA extraction protocol in Chapter 3 details the carefully handling of a basic protocol.

Before isolating the DNA, the bioinformatics tools for assembly and analysis should be considered. Understanding the workflow of a genome project and improvements provided by long-range sequencing approaches is discussed in Chapter 4. There are many ways to validate an assembly, and several quality control measures are described in the workflow presented. Beyond genome assembly, a significant challenge for any new genome project is annotating functional elements in the genome. There are a number of ways to annotate genes, and here, we present three chapters to assist with this process. Sometimes we seem to drown in the plethora of data generated as part of the genome project. Picking the best (maybe longest) transcripts from a list of expressed sequences and computer-generated forms is a daunting challenge that exceeds the capacity of any individual to manually annotate all the genes in an insect genome, but implementing specific software tools such as those described in Chapter 5 can produce a helpful set of “best” mRNAs to annotate a genome. An evaluation of different computational approaches to gene annotation is provided in this chapter, and a pipeline for generating automated gene predictions is developed and described. In addition to gene predictions, comparisons with other genomes can help us decide if the genome of interest is close to complete or lacking essential genes. Using an orthology-based framework, the BUSCO tool described in Chapter 6 allows us to assess sets of conserved single-copy genes across clades of insects at different phylogenetic levels. Once an assembly is considered “good enough” and automated gene predictions have been generated, a genome browser, hosted by the i5K workspace at the USDA national agricultural library, can be linked to web tools for manual curation of genes. This critical infrastructure greatly simplifies the organization of the research community most familiar with the biology of the organism under study and facilitates annotation of their favorite genes. Detailed protocols to extract data from this type of database along with quality control procedures which are critical to downstream analysis are presented in Chapter 7.

Access to the full-genome sequence provides a great deal more information than can be inferred from a transcriptome. One goal of many genome projects is to provide a resource that enables a translation of genome structure and content into studies of genome function. Interest in gene regulation begs the identification of promoters and enhancers. Three protocols in Chapters 8–10 describe FAIRE, RAMPAGE, and computational analysis of cis-regulatory modules in insects. Epigenetic modifications have also been implicated in patterns of gene regulation, and insect genomes display a wide range of methylation patterns that affect insect biology in a variety of ways. Chapters 11 and 12 focus on the generation of methylation patterns through whole-genome bisulfite sequencing and computational approaches to analyze these data.

Genome projects can be greatly enhanced by data from independent sources such as physical maps. Genome assemblies can be linked to chromosome-level architecture, and structural arrangements including large inversions can be described through Fluorescent In Situ Hybridization (FISH) techniques. The methods and application of FISH to examine genome structure in mosquito insects are described in Chapter 13.

Novel extensions of NGS sequencing approaches have opened up new avenues to study the microbial communities on and within insects. Many insects carry endosymbionts, the genomes of which are often revealed in the sequence data of the whole insect. In many cases, these microbial partners have profound impacts on the biology of their insect hosts as is the case with *Wolbachia* which can affect reproductive isolation and species divergence. A targeted approach to assay insect genomes (TEEseq) for the presence of endosymbionts is provided in Chapter 14.

The final chapter demonstrates the potential for powerful application of the data derived from insect genomes. In Chapter 15, the sequence of key developmental genes in mosquitos is used to develop an engineered yeast that constitutes a delivery system for lethal RNAi knockdowns. This method, which begins with a complete insect genome sequence, provides a generalizable model for RNAi systems that may find application in many pest and pathogen contexts.

As the price of next-generation sequencing goes down, the power of genomic approaches becomes affordable to virtually every biological inquiry. Along with the increased availability of these data is the need for powerful computational approaches to assemble and annotate insect genomes. The collection of protocols gathered herein is meant to provide an introduction to the workflows and bioinformatics tools available to researchers. These tools and approaches are rapidly evolving, so it is important to check for updates, improvements, and totally new tools that can perform the desired analysis.

We thank Michelle Coleman and Drs. Teresa Shippy and Sherry Miller for assistance in reviewing each chapter.

Manhattan, NY, USA
Notre Dame, IN, USA

Susan J. Brown
Michael E. Pfrender

Contents

<i>Preface</i>	<i>v</i>
<i>Contributors</i>	<i>ix</i>
1 Arthropod Genome Sequencing and Assembly Strategies	1
<i>Stephen Richards</i>	
2 Genome Size Estimation and Quantitative Cytogenetics in Insects	15
<i>J. Spencer Johnston, Angelina Bernardini, and Carl E. Hjelman</i>	
3 Isolation of High Molecular Weight DNA from Insects.....	27
<i>Susan J. Brown and Michelle Coleman</i>	
4 Long Range Sequencing and Validation of Insect Genome Assemblies.....	33
<i>Surya Saha</i>	
5 Integrated Modeling of Structural Genes Using MCuNovo	45
<i>Xiaolong Cao and Haobo Jiang</i>	
6 Using BUSCO to Assess Insect Genomic Resources.....	59
<i>Robert M. Waterhouse, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov</i>	
7 The GFF3toolkit: QC and Merge Pipeline for Genome Annotation.....	75
<i>Mei-Ju May Chen, Han Lin, Li-Mei Chiang, Christopher P. Childers, and Monica F. Poelchau</i>	
8 Using Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) to Identify Functional Regulatory DNA in Insect Genomes	89
<i>Daniel J. McKay</i>	
9 Using RAMPAGE to Identify and Annotate Promoters in Insect Genomes	99
<i>R. Taylor Raborn and Volker P. Brendel</i>	
10 CRM Discovery Beyond Model Insects.....	117
<i>Majid Kazemian and Marc S. Halfon</i>	
11 Whole-Genome Bisulfite Sequencing for the Methylation Analysis of Insect Genomes	141
<i>Fanny Gatzmann and Frank Lyko</i>	
12 Bioinformatic Analysis of Methylation Patterns Using Bisulfite Sequencing Data.....	157
<i>Jana Asselman</i>	
13 Physical Genome Mapping Using Fluorescence In Situ Hybridization with Mosquito Chromosomes	177
<i>Maria V. Sharakhova, Gleb N. Artemov, Vladimir A. Timoshevskiy, and Igor V. Sharakhov</i>	

14 Target-Enriched Endosymbiont Sequencing (TEEseq): A New High-Throughput Sequencing Approach Applied to the Comprehensive Characterization of Endosymbionts 195
Hannes Schuler, Jacqueline A. Lopez, Meredith M. Doellman, Glen R. Hood, Scott P. Egan, and Jeffrey L. Feder

15 Preparation and Use of a Yeast shRNA Delivery System for Gene Silencing in Mosquito Larvae 213
Keshava Mysore, Limb K. Hapairai, Na Wei, Jacob S. Realey, Nicholas D. Scheel, David W. Severson, and Molly Duman-Scheel

Index 233

Contributors

- GLEB N. ARTEMOV • *Department of Entomology and Fralin Life Science Institute, Virginia Tech, Blacksburg, VA, USA*
- JANA ASSELMAN • *Laboratory of Environmental Toxicology and Aquatic Ecology, Environmental Toxicology Unit (GhEnToxLab), Ghent University, Ghent, Belgium*
- ANGELINA BERNARDINI • *Interdisciplinary Program in Genetics, Texas A&M University, College Station, TX, USA*
- VOLKER P. BRENDEL • *Department of Biology, Indiana University, Bloomington, IN, USA; School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
- SUSAN J. BROWN • *Division of Biology, Kansas State University, Manhattan, KS, USA*
- XIAOLONG CAO • *Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK, USA*
- MEI-JU MAY CHEN • *Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan*
- LI-MEI CHIANG • *Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan*
- CHRISTOPHER P. CHILDERS • *Agricultural Research Service, National Agricultural Library, USDA, Beltsville, MD, USA*
- MICHELLE COLEMAN • *Division of Biology, Kansas State University, Manhattan, KS, USA*
- MEREDITH M. DOELLMAN • *Department of Biological Sciences, Galvin Life Sciences Building, University of Notre Dame, Notre Dame, IN, USA*
- MOLLY DUMAN-SCHEEL • *Department of Medical and Molecular Genetics, Indiana University School of Medicine, South Bend, IN, USA; Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA; Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA*
- SCOTT P. EGAN • *Department of BioSciences, Rice University, Houston, TX, USA*
- JEFFREY L. FEDER • *Department of Biological Sciences, Galvin Life Sciences Building, University of Notre Dame, Notre Dame, IN, USA*
- FANNY GATZMANN • *Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany*
- MARC S. HALFON • *Departments of Biochemistry, Biomedical Informatics, and Biological Sciences, University at Buffalo-State University of New York, Buffalo, NY, USA; NY State Center of Excellence in Bioinformatics and Life Sciences, Buffalo, NY, USA; Department of Molecular and Cellular Biology and Program in Cancer Genetics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA*
- LIMB K. HAPAIRAI • *Department of Medical and Molecular Genetics, Indiana University School of Medicine, South Bend, IN, USA; Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA*
- CARL E. HJELMEN • *Department of Entomology, Texas A&M University, College Station, TX, USA*
- GLEN R. HOOD • *Department of Biological Sciences, Galvin Life Sciences Building, University of Notre Dame, Notre Dame, IN, USA; Department of Biological Sciences, Wayne State University, Detroit, MI, USA*

- HAOBO JIANG • *Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK, USA*
- J. SPENCER JOHNSTON • *Department of Entomology, Texas A&M University, College Station, TX, USA*
- MAJID KAZEMIAN • *Departments of Biochemistry and Computer Science, Purdue University, West Lafayette, IN, USA*
- HAN LIN • *Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan*
- JACQUELINE A. LOPEZ • *Genomics and Bioinformatics Core Facility, University of Notre Dame, Notre Dame, IN, USA*
- FRANK LYKO • *Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany*
- DANIEL J. MCKAY • *Department of Biology, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; Department of Genetics, Integrative Program for Biological and Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*
- KESHAVA MYSORE • *Department of Medical and Molecular Genetics, Indiana University School of Medicine, South Bend, IN, USA; Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA*
- MONICA F. POELCHAU • *Agricultural Research Service, National Agricultural Library, USDA, Beltsville, MD, USA*
- R. TAYLOR RABORN • *Department of Biology, Indiana University, Bloomington, IN, USA*
- JACOB S. REALEY • *Department of Medical and Molecular Genetics, Indiana University School of Medicine, South Bend, IN, USA; Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA*
- STEPHEN RICHARDS • *Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA*
- SURYA SAHA • *Sol Genomics Network, Boyce Thompson Institute, Ithaca, NY, USA*
- NICHOLAS D. SCHEEL • *Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA; Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA*
- HANNES SCHULER • *Department of Biological Sciences, Galvin Life Sciences Building, University of Notre Dame, Notre Dame, IN, USA; Faculty of Science and Technology, Free University of Bozen-Bolzano, Bolzano, Italy*
- MATHIEU SEPPEY • *University of Geneva and Swiss Institute of Bioinformatics, Geneva, Switzerland*
- DAVID W. SEVERSON • *Department of Medical and Molecular Genetics, Indiana University School of Medicine, South Bend, IN, USA; Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA; Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA*
- IGOR V. SHARAKHOV • *Department of Entomology and Fralin Life Science Institute, Virginia Tech, Blacksburg, VA, USA*
- MARIA V. SHARAKHOVA • *Department of Entomology and Fralin Life Science Institute, Virginia Tech, Blacksburg, VA, USA*
- FELIPE A. SIMÃO • *University of Geneva and Swiss Institute of Bioinformatics, Geneva, Switzerland*
- VLADIMIR A. TIMOSHEVSKIY • *Department of Entomology and Fralin Life Science Institute, Virginia Tech, Blacksburg, VA, USA*

ROBERT M. WATERHOUSE • *Department of Ecology and Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, Lausanne, Switzerland*

NA WEI • *Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA; Department of Civil and Environmental Engineering and Earth Sciences, University of Notre Dame, Notre Dame, IN, USA*

EVGENY M. ZDOBNOV • *University of Geneva and Swiss Institute of Bioinformatics, Geneva, Switzerland*