Optimal High-Throughput Screening

This concise, self-contained, and cohesive book focuses on commonly used and recently developed methods for designing and analyzing high-throughput screening (HTS) experiments from a statistically sound basis. Combining ideas from biology, computing, and statistics, the author explains experimental designs and analytic methods that are amenable to rigorous analysis and interpretation of RNAi HTS experiments. The opening chapters are carefully presented to be accessible both to biologists with training only in basic statistics and to computational scientists and statisticians with basic biological knowledge. Biologists will see how new experimental designs and rudimentary data-handling strategies for RNAi HTS experiments can improve their results, whereas analysts will learn how to apply recently developed statistical methods to interpret HTS experiments.

Dr. Xiaohua Douglas Zhang is an associate director at Merck Research Laboratories. He has worked on data analysis for genome-wide RNAi research and microarrays in drug discovery and development for various diseases for many years. He has continuously developed novel analytic methods and experimental designs for quality control and hit selection in genome-scale RNAi research. Dr. Zhang and his colleagues have published many articles in various peer-reviewed journals, including *Cell Host & Microbe*, *Nucleic Acids Research*, *Bioinformatics*, *Genetic Epidemiology*, *Journal of Biological Chemistry*, *Pharmacogenomics*, *Genomics*, and *Journal of Biomolecular Screening*, among many others.

Optimal High-Throughput Screening

Practical Experimental Design and Data Analysis for Genome-Scale RNAi Research

Xiaohua Douglas Zhang
Merck Research Laboratories

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521734448

First published 2011

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data
Zhang, Xiaohua Douglas, 1966– author.
Optimal high-throughput screening : practical experimental design and data analysis for
genome-scale RNAi research / Xiaohua Douglas Zhang, Merck Research Laboratories.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-521-51771-3 (hardback) – ISBN 978-0-521-73444-8 (paperback)
1. High throughput screening (Drug development) 2. Small interfering RNA.
3. Experimental design. I. Title.
[DNLM: 1. High-Throughput Screening Assays – methods. 2. RNA Interference. 3. Research Design.
4. Statistics as Topic – methods. QV 778]
RS419.5.Z43 2011
615′.19–dc22 2010051114

ISBN 978-0-521-51771-3 Hardback
ISBN 978-0-521-73444-8 Paperback

Contents

Color plates follow page 110.

Preface

In 2000, scientists triumphantly announced they had deciphered the human genome, the blueprint for human life; in 2001, almost the entire human genome sequence became principally known. In 2003, the Human Genome Project was completed. By laying out in order the 3.2 billion units of our DNA, researchers sparked a firestorm of discovery and an explosion of genomic knowledge, which have been accompanied by rapidly emerging novel genomic technologies, including microarrays, whole-genome *single nucleotide polymorphism* (SNP) chips, *RNA interference* (RNAi) *high-throughput screening* (HTS), and so forth. All these have launched a new era – the genomic revolution era, which offers us boundless potential and great promise. Foremost are prospects in health, ranging from discovering cures for cancer to developing personalized medical products for individuals. The success in applying the "genomic revolution" to the discovery and development of new medical products largely depends on our ability to understand gene and gene interactions associated with drug response and disease. RNAi is a natural mechanism for gene silencing that can be harnessed to reveal information about gene function [48], leading to advances not only in drug target identification and validation, but also in the development of a potentially whole new class of therapeutic agents based on RNAi [24].

RNAi was first characterized as post-transcriptional gene silencing in petunia [109]. Later studies in *Caenorhabditis elegans* revealed that the interference with gene function was triggered by the presence of double-stranded RNA [48]. Exogenous delivery of double-stranded RNA was then developed as an experimental tool for functional genomics: first, in *C. elegans* and *Drosophila* and later, in mammalian cell culture systems, when it was discovered that delivery of short double-stranded RNA oligonucleotides triggers RNAi without inducing the interferon response [51]. This type of RNA oligonucleotide is thus called *small interfering RNA* (siRNA). The development of algorithms for siRNA design that produce a potent and selective knockdown of targeted genes has led to a great deal of interest in using siRNAs to elucidate gene function and identify novel targets for drug discovery. The importance

of RNAi was further recognized when the Nobel Prize in medicine and physiology was awarded to A. Fire and C. C. Mello in 2006 for their research in this field [48]. The application of genome-scale RNAi relies on the development of RNAi HTS technology.

RNAi HTS is broadly used in the identification of genes associated with specific biological phenotypes. This technology has been hailed as the second genomics wave, following the first genomics wave of gene expression microarray and single nucleotide polymorphism discovery platforms [101]. Before the emergence of RNAi HTS, compound HTS (which allows rapid screening of large collections of compounds consisting of small molecules) had been widely used in the pharmaceutical industry. As in any high-throughput platform, one of the most fundamental challenges in RNAi/compound HTS is gleaning biological significance from large volumes of data that rely on the development and adoption of appropriate statistical designs and analytic methods for quality control and hit selection [43].

Merck has applied extensive effort into RNAi research. It purchased Sirna Therapeutics for $1.1 billion in October 2006 [186] and has one of the largest labs for conducting genomewide RNAi research and compound HTS. Since early 2005, I have led data analysis in RNAi HTS projects at Merck and have continuously developed and adopted experimental designs and analytic methods for genome-scale RNAi research, including novel analytic methods for quality control and hit selection, which has allowed my colleagues and me to publish multiple articles on genome-scale RNAi research [28;45;86;161–180;182;183]. In 2005, I gave presentations on statistical methods for RNAi HTS at the Joint Statistical Meetings (Minneapolis, Minnesota) and the RNAi Meeting (Cold Spring Harbor, New York). Since then, I have given invited presentations and seminars at, among others, the 2006 International Conference on Bioinformatics and Computational Biology (Las Vegas, Nevada); 2006 Joint Statistical Meetings (Seattle, Washington); 2007 Seminars of Institute of Microbiology, Chinese Academy of Sciences (Beijing, China); 2007 International Conference of Bioinformatics (Hong Kong, China); 2007 Joint Statistical Meetings (Salt Lake City, Utah); 2008 Department of Statistics Seminar, Temple University (Philadelphia, Pennsylvania); 2008 RNAi and miRNA World Congress (Boston, Massachusetts); and 2009 World Pharmaceutical Congress (Philadelphia, Pennsylvania).

During my presentations, the following questions are usually asked by members of the audience: "As a statistician, I want to know about recently developed statistical methods and analytic tools in genome-scale RNAi research. Could you suggest a book in this field?" "As a biologist, although I have some knowledge of statistics, I do not have systematic training in this field. I am very interested in reading a book that introduces analytic tools in genome-scale RNAi research. Do you know of any book that describes the necessary and basic statistics knowledge in genome-scale RNAi research so that I can apply it to our experiments without knowing much details about statistics?" Or, "I am a graduate student. I am very interested

in data analysis in genome-scale RNAi research and am eager to work in this area in the near future whenever possible. Could you tell me which book will describe the necessary scientific knowledge and prepare me well for data analysis in this area?" Obviously, there is a demand for a self-contained and cohesive book about data analysis on genome-scale RNAi research. However, to my knowledge, such a book had yet to be written. The demands from the audiences and the need to promote genome-scale RNAi research propelled me to write such a book, in which I describe and present scientific knowledge and recently developed analytic methods and applications based on my experience in developing them and analyzing many genome-scale RNAi projects in the pharmaceutical industry.

Audience

In genome-scale RNAi research, it takes an ongoing dialog and a two-way flow of information and ideas between biologists and computational scientists, including statisticians, to develop experimental designs and analytic methods that are amenable to rigorous analysis and interpretation of RNAi HTS experiments. It has been recognized that biologists have an unfortunate tendency to "plug and play" with analytic methods without understanding the underlying principles, resulting in the misuse of otherwise effective strategies. Thus, at this time, most biologists depend on their computational colleagues for the development of data analysis methods, and most computational scientists depend on their biology colleagues to perform experiments that address important biological questions and to generate data [5]. Meanwhile, some people believe that, soon, if a scientist does not understand some statistics or rudimentary data-handling technologies, he or she may not be considered a true molecular biologist and thus will simply become a dinosaur [43]. On the other hand, to perform appropriate and effective data analysis, computational scientists need to know the details about recently developed methods and understand basic biological processes and technologies in HTS experiments.

Considering the needs of both biologists and computational scientists, this book has two major goals: 1) to help biologists who have limited training in statistics understand experiment designs, recently developed statistical methods, and rudimentary data-handling strategies for RNAi HTS experiments; and 2) to help computational scientists grasp recently developed statistical methods and common analytic tools and then be able to use them for analyzing data in HTS experiments. It is also suitable for graduate students (and perhaps undergraduate students) of biology or computational science who want to learn data analysis in HTS technologies. The first part of this book should be generally comprehensible to a biologist with training in basic statistics, as well as to a computational scientist with basic biological knowledge; the second part of this book should be comprehensible to a computational scientist with a master's degree or equivalent in statistics/biostatistics. The analytic methods presented in this book should also be suitable for biologists/chemists and

computational scientists working in any other HTS, including small-molecule HTS experiments.

Content and General Outline

This is a concise, self-contained, and cohesive book focusing on commonly used and recently developed methods for designing an HTS experiment from a statistically sound basis and for analyzing data from the experiment. The topics of this book reflect my personal experiences and biases in designing and analyzing HTS data. A significant portion of the book is built on material from articles that my colleagues and I have written, talks I have presented at multiple conferences, and my unpublished observations. Although I have tried to quote relevant literature, I may have missed some related references.

Chapter 1 presents an introduction to RNAi and HTS technologies and a description of a typical RNAi HTS project in the pharmaceutical industry. Chapter 2 provides experimental designs for genome-scale RNAi screens, which include siRNA designs, control designs, plate designs, designs for siRNA delivery and optimization of transfection, and sample size designs. Chapter 3 discusses how to display data in order to identify potential systematic errors, how to determine data transformation, and how to adjust for identified systematic errors. In Chapter 4, I present both biological processes and analytic methods for quality control and demonstrate how to apply them in HTS experiments. In an RNAi HTS, a primary goal is to select siRNAs with a desired size of inhibition or activation effect. An siRNA with a desired size of effect in an HTS screen is called a hit. The process of selecting hits is called hit selection. The analytic methods for hit selection in the screens without replicates differ from those with replicates. Therefore, I present them separately: without replicates in Chapter 5 and with replicates in Chapter 6. In Chapters 5 and 6, I explore classic analytic methods, including z-score method and t-test; describe recently developed methods, including robust methods, error control methods, and methods based on *strictly standardized mean difference* (SSMD) for hit selection; briefly introduce analytic methods for addressing off-target effects; and illustrate how to use them in RNAi HTS experiments. Sample size consideration for hit selection is also explored in Chapters 5 and 6. Chapters 1 through 6, which are presented in Part I of this book, are written so that a scientist with training in basic statistics can understand them. In each of these six chapters, I provide strategies on when and how to apply experimental designs and analytic methods in practical RNAi HTS experiments.

In contrast, Chapters 7 and 8, which are presented in Part II, describe and derive recently developed analytic methods from a solid statistical foundation and thus require the reader to have systematic training in statistics. Specifically, in Chapter 7, I present newly developed statistical methods for comparing groups, including contrast variable, c^+-probability, d^+-probability, *standardized mean of*

a contrast variable (SMCV), and their statistical estimation and inference; derive SMCV-based criteria for assessing the strength of group comparisons; and extend the concepts to the settings of multifactor *analysis of variance* (ANOVA). Chapter 7 builds a strong theoretical base for newly developed statistical methods for assessing the size of siRNA effects and for addressing off-target effects. In Chapter 8, I describe and derive newly developed statistical methods for assessing the size of siRNA effects, which includes SSMD and associated error-control methods, such as false discovery rate, false non-discovery rate, p-value, p^*-value, q-value, and q^*-value for hit selection in RNAi HTS experiments. In Chapter 8, I also elaborate on analytic methods adjusting for off-target effects. R functions for most analytic methods in this book will be formed into an R library and should be submitted to Bioconductor (www.bioconductor.org).

Acknowledgments

First, I would like to thank Senior Editor Lauren Cowles of Cambridge University Press for her continuous work in motivating and encouraging me to write this book; otherwise, I might not have considered this challenge. I am deeply indebted to many colleagues at Merck Research Laboratories for their input, assistance, and support. At the risk of omitting many, I would like to thank particularly Drs. Joseph F. Heyse, Keith A. Soper, and Daniel J. Holder for their full support and constructive comments; and Drs. Marc Ferrer, Amy S. Espeseth, Berta Strulovici, Raul Lacson, Eric N. Johnson, Ruojing Yang, David Ross, Francesca Santini, Shane D. Marine, Erica M. Stec, Namjin Chung, Weiqing Zhao, Richard A. Klinghoffer, Yaping Liu, Priya Kunapuli, Jayne Chin, Adam Gates, Jenny Tian, Anthony B. Kreamer, Richard R. Peltier, Michael J. Weber, Alexander McCampbell, Louis Locco, Eugen Buehler, David Stone, John Majercak, William J. Ray, and Michele Cleary for their excellent collaboration in RNAi HTS research at Merck. I would like to acknowledge Pamela Peterson and Brenda Holmes for proofreading help; my summer interns Pei-fen Kuan and Xiting Cindy Yang for their assistance; and Drs. Christopher Tong, Xiaoli Shirley Hou, Jason Liao, Yingxue Cathy Liu, Richard Raubertas, and Andy Liaw for their beneficial comments. I would like to thank two anonymous reviewers for their excellent comments and suggestions. I would also like to thank Troop 133 (Chalfont, PA) of Boy Scouts of America for their encouragement when I worked on this book during the 2009 summer camp. Finally and very importantly, I would like to thank my wife, Xiaohua Shu, and my son, Zhaozhi Zek Zhang, for their strong support and great patience during the writing of this book. Without the invaluable input and support of the many professionals I have mentioned and of my family, this book would not have become a reality.

Acronyms and Abbreviations

3′UTR	3′ untranslated region
ANOVA	analysis of variance
AUE	approximately unbiased estimate
CI	confidence interval
dsRNA	double-stranded RNA
esiRNA	endoribonuclease-derived siRNA
FDR	false discovery rate
FNDR	false non-discovery rate
FNL	false-negative level
FNR	false-negative rate
FPL	false-positive level
FPR	false-positive rate
HCV	hepatitis C virus
HTS	high-throughput screening
MAD	median absolute deviation
miRNA	microRNA
MLE	maximum likelihood estimate
MM	method of moment
mRNA	messenger RNA
MSE	mean squared error
pnc	proportional noncentral
QC	quality control
RISC	RNA-induced silencing complexes
RNAi	RNA interference
SD	standard deviation
shRNA	short hairpin RNA
siRNA	small interfering RNA

SMCV	standardized mean of a contrast variable
SMLC	standardized mean of a linear combination of random variables
SSMD	strictly standardized mean difference
UMVUE	uniformly minimal variance unbiased estimate
w.r.t.	with respect to