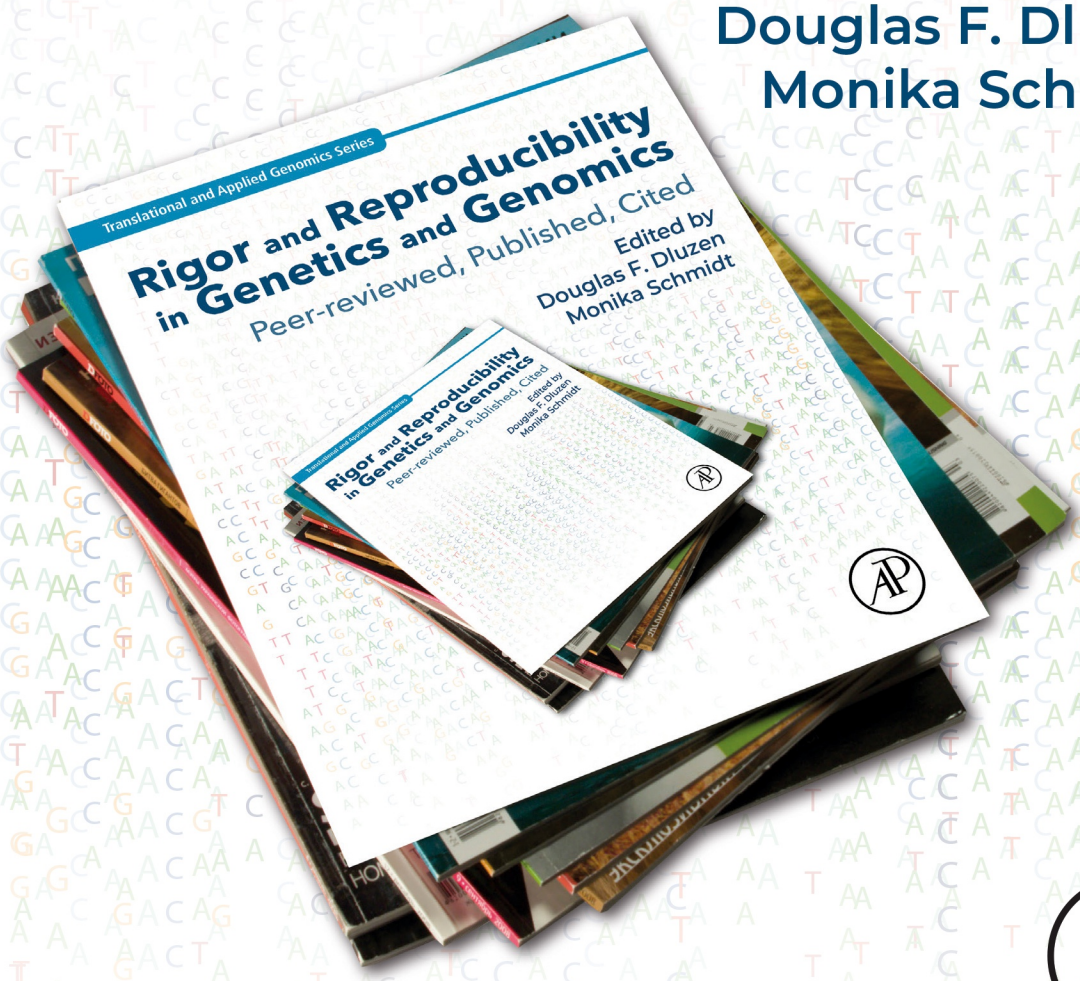
Translational and Applied Genomics Series

Rigor and Reproducibility in Genetics and Genomics

Peer-reviewed, Published, Cited

Edited by
Douglas F. Dluzen
Monika Schmidt



Rigor and Reproducibility in Genetics and Genomics

This page intentionally left blank

Translational and Applied Genomics

Rigor and Reproducibility in Genetics and Genomics

Peer-reviewed, Published, Cited

Edited by

Series Editor

George P. Patrinos

Professor, Department of Pharmacy, University of Patras, School of Health Sciences, Patras, Greece; United Arab Emirates University, College of Medicine and Health Sciences, Department of Pathology, Al-Ain, UAE; United Arab Emirates University, Zayed Center of Health Sciences, Al-Ain, UAE; Erasmus University Medical Center, School of Medicine and Health Sciences, Department of Pathology – Bioinformatics Unit, Rotterdam, The Netherlands

Series Volume Editors

Douglas F. Dluzen

Visiting Professor of Biology, Morgan State University, Baltimore, MD, United States; Office of Graduate Biomedical Education, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Monika H.M. Schmidt

Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada

ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

ISBN 978-0-12-817218-6

Publisher: Stacy Masucci
Acquisitions Editor: Peter B. Linsley
Editorial Project Manager: Susan E. Ikeda
Production Project Manager: Jayadivya Saiprasad
Cover Designer: Mark Rogers

Typeset by STRAIVE, India

Contents

Contributors xv
Preface xix

SECTION 1 Introduction

CHAPTER 1 Rigor and reproducibility in genetic research and the effects on scientific reporting and public discourse 3
Monika H.M. Schmidt and Douglas F. Dluzen

Introduction 3
What is the rigor and reproducibility crisis? 5
The issue of waning public trust in scientific research 7
What are the contributing factors to the reproducibility crisis? 9
The societal importance of open science 12
Conclusions 15
References 17

CHAPTER 2 Unveiling the hidden curriculum: Developing rigor and reproducibility values through teaching and mentorship 23
Marina E. Tourlakis

Introduction 23
In the classroom 24
In practice 27
In advising 34
In the research lab 35
In practice 37
Conclusion and future work 42
References 43

SECTION 2 Genotyping

CHAPTER 3 Genome-wide association studies (GWAS): What are they, when to use them? 51
Fan Wang

Introduction and history 51
Study design and statistical methods 52
Frequentist vs Bayesian modeling 52
GWAS for case and control association testing 54
GWAS for quantitative traits 56

Quality control for GWAS 58
Technical QC 59
Genetic QC 59
Concluding note 60
Using R and PLINK for GWAS 61
R primer 61
Plink primer 61
Plink data formats 64
Genome-wide example 67
References 76

CHAPTER 4 GWAS in the learning environment 83
Amy L. Stark

Introduction 83
Theoretical background for designing and interpreting GWAS 83
What is GWAS? 83
What is a GWAS stated simply? 84
GWAS evaluation 85
Populations included in the GWAS 86
Population diversity 86
Population size 87
Number of tests in a GWAS 87
The trait in question: The phenotype 89
Summary 90
Reference 90

CHAPTER 5 Polygenic risk scores and comparative genomics: Best practices and statistical considerations 91
Sally I-Chun Kuo and Fazil Aliev

Genome-wide association studies 91
The polygenic scoring approach 92
Preparing data for polygenic score calculation 93
Software programs for PRS calculation 97
Sample script for calculating PRS using PRS-CS 99
Challenges associated with creating predictive PRS and interpreting PRS results 99
Conclusion 101
Appendix A 101
Example script for creating polygenic scores 101
Using single R script to prepare and send all commands to the system to create PRS-CS score 104
References 108

CHAPTER 6 Sequence analysis and genomics in the classroom 115
Rebecca C. Burgess, Rivka Glaser, and Kimberly Pause Tucker

Crowd-sourced undergraduate research 115
The Genomics Education Partnership 116
SEA-PHAGES 119
Standalone DNA sequencing activities 120
16S rRNA bacterial genotyping/identification activity 120
Building and interpreting molecular phylogenies 123
Bitter tasting ability (PTC) genotyping activity 127
Other resources for DNA sequencing and analysis activities 132
National Center for Case Study Teaching in Science 132
Association for Biology Laboratory Education 132
Genetics Society of America 132
HHMI biointeractive 133
QUBES 133
GCAT-SEEK 133
Bio-rad cloning and sequencing explorer series 133
References 134

CHAPTER 7 Classroom to career: Implementation considerations for engaging students with meaningful DNA sequencing learning opportunities 137
Charles Wray

Introduction 137
DNA sequencing at the high school level 138
Essential content and learning goals 138
Technology considerations for high schools 141
DNA sequencing at the undergraduate level 142
Engaging and exciting students 144
Promoting research careers 145
Skills and career training 147
Future considerations 149
References 149

SECTION 3 Next-generation sequencing & gene expression

CHAPTER 8 Review of gene expression using microarray and RNA-seq 159
Ana B. Villaseñor-Altamirano, Yalbi Itzel Balderas-Martínez, and Alejandra Medina-Rivera

Introduction 159
High-throughput techniques to assess gene expression 161

Microarrays—What are they, and how are they conducted?....161
Sequencing....163
How does one overcome read length challenges?....165
Applications....166
Gene expression profiling....166
Splicing detection....168
Expression quantitative trait loci assessment....170
Single-cell RNA-seq....172
Public databases....173
Databases....173
Reproducibility across studies....176
Key point reproducibility, replicability, robustness, generalizability....176
Batch effect....177
Metaanalysis....177
Conclusion and remarks....178
Acknowledgments....178
References....179

CHAPTER 9 Guidelines and important considerations for 'omics-level studies....189
Francesca Luca and Athma A. Pai

Overview of the RNA-sequencing experimental protocol....190
Study design considerations—Definitions....192
RNA quality....192
Confounders....193
Replicates....193
Sequencing depth....195
Read length and type....196
Analysis of RNA-seq data....197
Quality control....197
Mapping of reads....197
Quantifying gene and isoform expression....199
Quantifying absolute mRNA abundance....200
Comparing gene expression between groups....201
Studying the genetic determinants of gene expression....202
Analysis of alternative splicing events....205
Summary....206
References....207

CHAPTER 10 Rigor and reproducibility of RNA sequencing analyses 211
Dominik Buschmann, Tom Driedonks, Yiyao Huang, Juan Pablo Tosar, Andrey Turchinovich, and Kenneth W. Witwer

Rigor and reproducibility of RNA sequencing analyses 211
Introduction to RNA sequencing 211
Extracellular RNA: A case study for the challenges of low-input RNA-Seq 212
Experimental design and preanalytical variables 213
The measurement process 213
Sample preservation: What is the original state? 214
Contamination introduced before or during library preparation 215
Quality control 216
Experimental design and validation 217
The impact of library preparation methods 218
RNA and cDNA fragmentation 218
Reverse transcription 220
Adapter attachment 222
Library amplification and unique molecular identifiers 223
Impact of size selection on RNA fragment distribution and detected RNA biotypes 223
Data analysis 225
Preprocessing 225
Read mapping 226
Overlapping annotations and database quality 227
Read quantification 228
Data quality control 228
Normalization and differential expression analysis 229
Functional analysis of RNA-Seq data 230
Conclusions 230
References 231

CHAPTER 11 Validation of gene expression by quantitative PCR 247
Arundhati Das, Debojyoti Das, and Amaresh C. Panda

Introduction 247
Materials 248
RNA extraction 248
Assessment of RNA 248
cDNA synthesis and qPCR 248
Method 249
Total RNA isolation 249

Analysis of RNA quantity, purity, and integrity....250
Reverse transcription of RNA....250
Designing RNA-specific primers....252
Validation of expression by quantitative (q)PCR....254
Technical notes....255
Acknowledgments....256
Conflict of interest....256
References....256

SECTION 4 Epigenetic analyses

CHAPTER 12 Best practices for epigenome-wide DNA modification data collection and analysis....261
Joseph Kochmanski and Alison I. Bernstein

Introduction....261
DNA modifications....261
DNA modifications in health and disease....262
Methods for detection of DNA modifications....264
Bisulfite conversion methods....264
Alternatives to BS-based methods....265
Challenges to reproducibility in DNA modification EWAS research....265
Biology....265
Methodology....267
Statistics....268
Experimental planning & reporting: Methods, code, and data....272
Conclusion....274
References....274

CHAPTER 13 Best practices for the ATAC-seq assay and its data analysis....285
Haibo Liu, Rui Li, Kai Hu, Jianhong Ou, Magnolia Pak, Michael R. Green, and Lihua Julie Zhu

An overview of ATAC-seq....285
Generating high-quality ATAC-seq data....287
Experimental design....287
Nuclei preparation and quality control....288
Tagmentation, PCR amplification, and quality control....289
Sequencing....290
Analyzing ATAC-seq data: From stringent data quality control to comprehensive data mining....291
Raw read quality control and preprocessing....291
Alignment, postalignment processing, and quality control....291
Peak calling....293

Differential peak analysis....296
Annotation and functional analysis....297
Visualization....297
Nucleosome positioning....298
TF occupancy inference....299
Motif mapping....303
Differential TF binding activity analysis....303
Reconstruction of gene regulatory network....304
Summary....305
Acknowledgments....309
References....309

CHAPTER 14 Best practices for ChIP-seq and its data analysis....319
Huayun Hou, Matthew Hudson, and Minggao Liang

Introduction....319
Crucial considerations for a rigorous ChIP-seq experiment....320
To fix or not to fix?....320
Nuclear isolation for ChIP-seq....321
Fragmenting chromatin for ChIP-seq....321
Single-cell and low-input ChIP-seq....322
Appropriate controls in ChIP-seq....322
Antibody incubation, chromatin washing, and elution....323
Analysis of ChIP-eluted chromatin....323
Library preparation....323
Sequencing a ChIP-seq experiment....325
Sequencing read preprocessing and alignment....326
Peak calling....326
Quality control (QC)....327
Visualization of ChIP-seq data....328
Peak annotation and functional enrichment analysis....329
Differential binding analysis....331
Motif analysis....332
Key take-away....333
Conclusion....334
References....335

CHAPTER 15 A practical guide for essential analyses of Hi-C data....343
Yu Liu and Erica M. Hildebrand

A brief summary of Hi-C and an example analysis pipeline....343
Hi-C data processing and quality control....345
Visualization of Hi-C data....346
Scaling plots and chromosome folding....348

Compartment analysis....349
Insulation and TAD boundaries....351
Dot calling and CTCF-CTCF loops....352
Integration of Hi-C data with ChIP-seq and RNA-seq data....354
Data use....357
Computational resources....357
Acknowledgments....357
References....357

CHAPTER 16 Epigenetics in the classroom....363
Khadijah Makky

Epigenetics in undergraduate biology curriculum: Why, when, where, and how....363
How is the chapter organized?....364
Where is it appropriate to introduce epigenetics in the biology curriculum?....364
How to design your epigenetics unit....366
Approach to teaching epigenetics using high-impact practices....368
Learning outcome 1: Building the epigenetics foundation knowledge....368
Regulation of gene expression and epigenetics....369
Epigenetic marks and chromatin conformation....370
How does the chromatin conformation change?....370
Epigenetic marks....370
Learning outcome 2: Application using basic knowledge to critically understand many epigenetic phenomena....373
Genomic imprinting—Using compare/contrast....373
X-inactivation—Using concept maps....373
Summarizing information—Using case studies....375
Learning outcome 3: Integration-connecting the knowledge from this unit to the realms of life....377
Epigenetics and cancer....377
Culminating case study....377
Case presentation in the classroom....378
Epigenetics and human behavior: Nature versus nurture....384
Conclusion and future considerations....391
References....391

SECTION 5 Gene editing technologies

CHAPTER 17 Genome editing technologies....397
Dana Vera Foss and Alexis Leigh Norris

Introduction....397
Zinc finger nucleases....399
Background....399

Applications 400
Future directions 402
Transcriptional activator-like effector nucleases 402
What they are 402
Applications 404
Future directions 405
CRISPR-Cas systems 405
Background 405
Applications 406
Future directions 407
Delivery of genome editing systems 407
Background 407
Methods 408
Future directions 410
Gene drive systems 411
What they are 411
How they are used 412
Future directions 412
Unintended genomic alterations 413
Background 413
Methods 414
Future directions 416
Conclusion 416
References 417

CHAPTER 18 Genetic modification of mice using CRISPR-Cas9: Best practices and practical concepts explained 425
Vishnu Hosur, Benjamin E. Low, and Michael V. Wiles

Genetically engineered mouse models of human disease 425
International Knockout Mouse Consortium (IKMC) 426
Cancer 426
Alzheimer's disease (AD) 427
COVID-19 427
Generating mouse models using CRISPR-Cas9 427
Methodology 428
Genetic diversity in mice 430
Reagent delivery to the mouse zygote 431
Guide design 431
Key considerations for generating KO alleles 434
Verification of KO alleles 436
Verification of dropout alleles 436
Key considerations for generating small knock-in alleles 438
Verification of small knock-in alleles 438

Key considerations for generating large knock-in alleles 440
Verification of large knock-in alleles 440
Practical methods 441
Mosaicism 442
Unintended consequences 443
Off-targeting events can occur but are circumventable 443
Unintended on-target effects are likely but preventable 443
Conclusions and future perspective 444
Cas9 variants 444
Delivery of CRISPR-Cas9 components 444
Base editing 445
Prime editing 445
Acknowledgments 446
References 446

CHAPTER 19 CRISPR classroom activities and case studies 453
TyAnna L. Lovato and Richard M. Cripps

Importance of course-based undergraduate research experiences 453
Importance of CRISPR as a teaching tool 453
Approaches to teaching CRISPR 454
Setting the stage 454
Strategies for short time frames in bacteria 455
16week teaching strategies 456
General considerations for developing CRISPR in the classroom 456
Classroom activities at the University of New Mexico 457
Designing and cloning CRISPR targets in Drosophila 457
Recent innovations at the University of New Mexico 460
Considerations of rigor and reproducibility in CRISPR classes 462
Conclusions 464
Appendix: Detailed methods used to create guide RNA plasmids for use in Drosophila 464
Exercise 1: Annealing oligonucleotides to generate short dsDNA inserts 464
Exercise 2: Phosphorylating your dsDNA, and ligating into pBFv-U6.2 465
Exercise 3: Transformation of ligation products into *E. coli* 465
Exercise 4: Picking colonies, minipreps, and sequencing 466
Exercise 5: Sequencing clean-up 467
Exercise 6: Analysis of sequences and transformation of successful clones 468
Acknowledgments 470
References 470

Index 473

Contributors

Fazil Aliev
Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers Behavioral and Health Sciences, Piscataway, NJ, United States

Yalbi Itzel Balderas-Martínez
Instituto Nacional de Enfermedades Respiratorias Ismael Cosio Villegas, Laboratorio de Biología Computacional, Mexico City, Mexico

Alison I. Bernstein
Department of Pharmacology and Toxicology; Environmental and Occupational Health Science Institute, Rutgers University, Piscataway, NJ, United States

Rebecca C. Burgess
Stevenson University, Owings Mills, MD, United States

Dominik Buschmann
Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Richard M. Cripps
Department of Biology, San Diego State University, San Diego, CA, United States

Arundhati Das
Institute of Life Sciences, Bhubaneswar, Odisha, India

Debojyoti Das
Institute of Life Sciences, Bhubaneswar, Odisha, India

Douglas F. Dluzen
Office of Graduate Biomedical Education, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Tom Driedonks
Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Dana Vera Foss
Wilson Lab, University of California Berkeley, Berkeley, CA, United States

Rivka Glaser
Stevenson University, Owings Mills, MD, United States

Michael R. Green
Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School, Worcester, MA, United States

Erica M. Hildebrand
Department of Systems Biology, University of Massachusetts Chan Medical School, Worcester, MA, United States

Vishnu Hosur
The Jackson Laboratory, Bar Harbor, ME, United States

Huayun Hou
Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada

Kai Hu
Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School, Worcester, MA, United States

Yiyao Huang
Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, MD, United States; Department of Laboratory Medicine, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, China

Matthew Hudson
Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada

Joseph Kochmanski
Rancho BioSciences, San Diego, CA; Department of Translational Neuroscience, Michigan State University, East Lansing, MI, United States

Sally I-Chun Kuo
Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers Behavioral and Health Sciences, Piscataway, NJ, United States

Rui Li
Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School, Worcester, MA, United States

Minggao Liang
Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada

Haibo Liu
Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School, Worcester, MA, United States

Yu Liu
Department of Systems Biology, University of Massachusetts Chan Medical School, Worcester, MA, United States

TyAnna L. Lovato
Department of Biology, University of New Mexico, Albuquerque, NM, United States

Benjamin E. Low
The Jackson Laboratory, Bar Harbor, ME, United States

Francesca Luca
Center for Molecular Medicine and Genetics; Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, United States; Department of Biology, University of Rome “Tor Vergata”, Rome, Italy

Khadijah Makky
Department of Biomedical Sciences, Marquette University, Milwaukee, WI, United States

Alejandra Medina-Rivera
Laboratorio Internacional de Investigación sobre el Genoma Humano, UNAM, Querétaro, Mexico

Alexis Leigh Norris
Food and Drug Administration, Bioinformatician, Center for Veterinary Medicine, Rockville, MD, United States

Jianhong Ou
Department of Cell Biology, Duke University School of Medicine, Duke Regeneration Center, Duke University, Durham, NC, United States

Athma A. Pai
RNA Therapeutics Institute, University of Massachusetts Chan Medical School, Worcester, MA, United States

Magnolia Pak
Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School, Worcester, MA, United States

Amaresh C. Panda
Institute of Life Sciences, Bhubaneswar, Odisha, India

Monika H.M. Schmidt
Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada

Amy L. Stark
University of Notre Dame, Notre Dame, IN, United States

Juan Pablo Tosar
Nuclear Research Center, School of Science, Universidad de la República; Functional Genomics Laboratory, Institut Pasteur de Montevideo, Montevideo, Uruguay

Marina E. Tourlakis
Biology Department, University of the Fraser Valley, Abbotsford, BC, Canada

Kimberly Pause Tucker
Stevenson University, Owings Mills, MD, United States

Andrey Turchinovich
Division of Cancer Genome Research, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK); Heidelberg Biolabs GmbH, Heidelberg, Germany

Ana B. Villaseñor-Altamirano
Laboratorio Internacional de Investigación sobre el Genoma Humano, UNAM, Querétaro, Mexico

Fan Wang
University of Chicago, Center for Translational Data Science, Chicago, IL, United States

Michael V. Wiles
The Jackson Laboratory, Bar Harbor, ME, United States

Kenneth W. Witwer
Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Charles Wray
The Jackson Laboratory, Genomic Education, Bar Harbor, ME, United States

Lihua Julie Zhu
Department of Molecular, Cell and Cancer Biology; Department of Molecular Medicine, Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, United States

Preface

The growth of scientific knowledge is rarely linear. Historically, the pace of discoveries was sufficiently gradual to permit revision of proposed theories in a timely manner or, at least, without massive investment of resources. In recent years, the landscape of how genetic and genomic research is conducted has rapidly changed with the advent of the age of computing. In silico research and computational experiments complementing traditional at-the-bench research now represent a significant portion of newly published research, and it is published at an astonishing pace. Moreover, new computational methods and their related bench techniques are continuously under development, discussed at conferences, and, increasingly, promoted on preprint servers for public consumption.

Preprint servers in and of themselves present a new challenge to the field of biomedical science: although these servers increase accessibility of scientific research, particularly for the public, they also provide opportunity for nonpeer-reviewed content to be widely disseminated, irrespective of the quality or reproducibility of data or methods presented. Issues relating to incomplete or incorrect reporting of such findings by news outlets and other nonexpert media personalities are merely one consideration of the importance of rigorous, reproducible methods and reporting standards. For genomics researchers, rigorous methods and detailed documentation pertaining to computational tools are absolutely crucial at all times: during critical evaluation of preprint publications by fellow scientists, during peer review, and long into the future, should another researcher choose to adopt the same computational method or tool in their work.

The rapid pace of new developments in genetics and genomics comes with an additional caveat: It makes educational textbooks, like this one, seemingly out-of-date by publication. Yet, providing cutting-edge methods is not the goal of this book; this book is concerned with providing guidelines and principles for conducting reproducible, high-quality genomic research. It is neither a reference manual nor an encyclopedia of methods, as the staggering number of computational tools and in silico techniques querying ever more complex ideas cannot be captured within the physical constraints of a book, or even an anthology of books!

This (e-)book seeks to provide one of the first compilations of genomic techniques with a focus on addressing the reproducibility crisis currently faced by biomedical research. Admittedly, the mountain to climb in this regard is enormous and will require coordinated efforts from granting bodies, publishers, and researchers themselves. Nonetheless, it begins—as with all systemic changes in a society—with educating the newest members of the genetics and genomics research community: trainees, early career investigators, and lecturers teaching this material. This is our intended audience, and the contents of each chapter will reflect this angle.

Rigor and Reproducibility in Genetics and Genomics is chiefly concerned with laying a foundation of basic "dry lab" methodologies and providing thoughtful examples of how to pivot to new approaches while still upholding rigorous scientific practice to produce reproducible outcomes. This book originated as an Invited Session at the 2017 American Society of Human Genetics Annual Meeting in Orlando, Florida. We attempted to include as many topics as we felt this book could reasonably discuss, and selected methods and computational research areas that are rapidly growing or already widely adopted. Our authorship is reflective of the diversity and global nature of genetic and genomic researchers, a key principle we kept in mind during the recruitment phase for this book.

We assume that most readers have a basic understanding of genetics and genomics, but have nonetheless attempted to include one or more review chapters in each section (see Chapters 3, 8, 12, and 17) providing a brief overview of the techniques to be discussed in subsequent chapters. Where possible, we have included teaching resource chapters written by expert undergraduate educators (Chapters 2, 4, 6, 7, 16, and 19). The intervening chapters provide relevant examples and protocols for some of the most au courant approaches in genetic and genomic research. These chapters also highlight the merits and drawbacks to any particular methodology or computational tool, as well as key considerations when developing a research pipeline using the technique under examination. This book will put readers on solid footing when looking to apply the discussed genomic techniques to their work.

The greatest thanks and acknowledgments are owed to each of the chapter authors: for their time, patience, and expert contributions. The COVID-19 pandemic extended the project timeline on the development of this book in unimaginable ways. The first year (or more) of the pandemic paused facets of research and complicated everyone's personal lives, yet our authors pushed through—this speaks volumes about the importance they placed on the written contents between these covers. Many of these chapters were coauthored by doctoral trainees or postdoctoral researchers, who are often at the leading edge of research and developing improved research methods. This book was written by them with you, the reader, at the forefront.

We would also like to thank the editing team at Elsevier, in particular Peter Linsley, who recognized the importance of this topic and approached us with this opportunity to educate. As well, our senior editorial project managers, Susan Ikeda and Kristi Anderson, who worked tirelessly to keep this project moving toward completion. In particular, a special thank you to Susan for her patient understanding and warm encouragement as we faced various editing hurdles.

Finally, a huge thanks to our families, who were considerate in their time and patience as we worked on this book at all hours of the day (and night). We have each navigated the wonderful arrival of two children apiece, further motivating our desire to set up young trainees with a new resource that can serve as a guide during their research careers, establishing a brighter future for biomedical research.

We hope you find this book knowledge-dense and resource-intensive in a directly applicable sense, and wish you the best in your genetic and genomic research journey!

Douglas F. Dluzen
Monika H.M. Schmidt

SECTION

Introduction 1

This page intentionally left blank

CHAPTER

1

Rigor and reproducibility in genetic research and the effects on scientific reporting and public discourse

Monika H.M. Schmidt[a] **and Douglas F. Dluzen**[b]

[a]*Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada,* [b]*Office of Graduate Biomedical Education, Johns Hopkins University School of Medicine, Baltimore, MD, United States*

Introduction

The scientific method has been practiced by humankind throughout our evolution, as we engaged in *trial and error* and worked toward finding better ways to survive and thrive. At its most basic, the scientific method requires the observer to integrate known information about a situation and process through influencing factors as the observer puts into motion a plan to obtain a desired outcome. Whether or not one is scientifically trained, everyone has practiced this form of logical thinking at some point in their lives. For example, imagine coming home after a long day at work, sitting down in a favorite chair or couch, and turning on the television, but the television does not turn on. You hit the power button on the remote again—nothing. Disbelief and frustration might begin. You must now work through the different factors inhibiting your relaxation and enjoyment.

> Anything blocking the signal path between remote and television? No? Check.
> Power to the television? Yes. Check.
> Power to the living space? Yes. Check. (And likely integrated into consideration already).
> Batteries in the remote dead? Swap and replace—then retest. Bingo!

The scientific method is a problem-solving tool designed to give us a certain degree of confidence when we finally obtain a result, whether it was predicted or not. The conclusions drawn from even the simplest of experiments are only as strong as the weakest point in the underlying approach to generating, collecting, and analyzing the data from that approach or experimental design. The same is true in genetic and genomic research.

Strong experimental designs accounting for confounding variables are needed to untangle the complex factors that may influence the outcomes of any given genetic or genomic study. This is especially true for analyses that incorporate information from large population data sets. This chapter and those that follow in this textbook resource examine some of the ways in which we can structure the most

Rigor and Reproducibility in Genetics and Genomics. https://doi.org/10.1016/B978-0-12-817218-6.00012-7

widely used experimental approaches in genetic and genomic research to increase the confidence, replicability, and applicability of the results.

Historically speaking, scientific "proof" used to require that one could demonstrate a scientific phenomenon in front of other scientists. There would be documentation of experiments with written word, and illustrations came later to allow readers to imagine being in the room, observing the experiment, and thereby accelerating the pace of dissemination of scientific research and its outcomes. While the general public may have often been invited to these discussions, debates, and lectures, they were not usually involved in the interpretation and advancement of the work. That has changed in the last few decades as news media, patient advocates, and those interested in the societal impact of publicly or privately funded science have become a necessary and essential component of the discussion of scientific advancement. This is especially true when we consider how the knowledge generated in the laboratory or clinic is *applied* in daily life.

The scientific discourse and review that validate new research results are tiered:

1. The first tier—the choice of methodology and the approaches taken by the authors of a given work and their collaborators.
2. The second tier—the review by the research community (grant review panels, conferences, and journal manuscript peer reviewers).
3. The third tier—feedback from the wider research community once a manuscript is submitted to a preprint service and/or formally accepted for publication in a peer-reviewed journal.
4. Fourth tier—delivery of research findings to the broader public where they may interact with the data, interpreting the applicability of the results to public policy or healthcare practices, or even providing the foundation to answer subsequent questions unearthed in the original study.

Breakdowns anywhere within or between these tiers have historically contributed to the publication of results that may have been misinterpreted, overly conflated, falsified, or fabricated, and have allowed methodologies inappropriately chosen to give a false sense of confidence with a study's results. *Research in many areas has gently shifted from a culture of "show me" to "trust me"—a defining reason for the need to ensure reproducibility of scientific works*.

In the field of genetics and genomics, advancing technology and statistical methods can be so diverse and complex that it is difficult to describe them even to a technical audience. Peer reviewers and journal editors are required to review enormous volumes of submissions and to have a wide breadth of expertise, without having sufficient information (or time) to do their jobs thoroughly, thereby inadvertently permitting problematic research to slip through the peer review processes. The myriad reasons underlying this problem relate to funding challenges and a *publish or perish* attitude that underlies much of biomedical research—but some of these systemic issues are beyond the scope of this book.

There are numerous other concerns in the scientific community that can contribute to published research that is not methodologically sound or able to be reproduced by other laboratories. In the past two decades, the subfield of meta-research has emerged, in which statisticians, researchers, and clinicians have examined the nature of the scientific method itself within biomedical and genetic research in order to identify key factors that influence the reliability and replicability of peer-reviewed science [1,2]. Meta-research has identified a possible rigor and reproducibility crisis in peer review and publishing processes as more and more manuscripts are published containing science that cannot be replicated and/or using inappropriate approaches for the given context. Further, due to the aforementioned *publish or perish* culture that is particularly prevalent in competitive research environments, combined

with digitally rendered data figures, the publication of difficult-to-detect but completely falsified data has had a marked uptick. A collaborative effort by researchers to identify and report such falsifications is necessary—an excellent example of image forensics is the work of Dr. Elisabeth Bik (Twitter: @MicrobiomDigest) [3], discussed in further detail here.

This chapter is dedicated to introducing the historical context of this potential crisis (which some argue is also an *opportunity* for change), identifying systemic factors that may have contributed to the lack of replication within scientific studies and reproducibility by other groups, and suggestions for geneticists on key steps to improve upon communicating with the public on these issues.

Key point: reproducibility versus replicability

The terms reproducibility and replicability are used in this chapter and throughout this book. The difference between these terms is subtle, so much so that these terms are often used interchangeably—albeit incorrectly. Toward fostering rigorous attention to all details in scientific research, including language, we suggest that the definitions as outlined by the National Academy of Sciences in their 2019 book Reproducibility and Replicability in Science [4] be adopted across scientific communities. Thus:

Reproducibility is the ability to consistently obtain the same results using identical input data or reagents examined via the same experimental conditions and analyses.

Replicability is the ability to obtain consistent (but not necessarily identical) results when using different input data or reagents with the goal of answering the same scientific question.

If this seems confusing, consider an analogy involving baking a chocolate chip cookie: A reproducible batch of cookies will use identical ingredients (the same flour, same butter, same chocolate chips, same sugar, same water) and identical apparati (the same oven with the same cookie sheet) and identical baking conditions (same bake time and temperature). Assuming the recipe instructions are clear and detailed (no "add a thimbleful of baking powder") and that the ingredients are pure (the flour should not have any contaminants in it), the baker will likely be able to consistently produce the same delicious batch of chocolate chip cookies. A replicable batch of cookies will strive to consistently achieve delicious golden-on-the-outside and gooey-in-the-centre chocolate chip cookies, but may use ingredients produced by different companies, apparati with slight differences (air bake sheets versus plain aluminum bake sheets, for example) and may even follow slightly different instructions. Presumably though, with the same question of achieving the aforementioned cookie, a replicable chocolate chip cookie (not an oatmeal cookie) will be achieved.

What is the rigor and reproducibility crisis?

Rigorous and reproducible research practices are the bedrock of scientific advancement. One of the more thorough and recent reexaminations of the scientific method began in 2005 with an essay written by Dr. John Ioannidis. Dr. Ioannidis made a claim with far-reaching implications: that much of the published research findings were false [5]. He discussed that most studies were too small, underpowered, and/or included biases in study design, implementation, data collection and/or analysis, interpretation, and reporting. Ioannidis argued, "most research questions are addressed by many teams, and it is misleading to emphasize the statistically significant findings of any single team. What matters is the totality of the evidence. Diminishing bias through enhanced research standards and curtailing of prejudices may also help."

Ioannidis' work, and that of others, initiated a much-needed conversation identifying the qualities of a successful research study. Most scientific disciplines have now re-examined standard research protocols and practices, and found varying degrees of replication of prior studies. For example, the Reproducibility Project: Cancer Biology replicated 50 experiments from 23 high-impact cancer-related research papers [6]. The study investigators replicated less than half of the experiments that provided

positive results, but nearly 80% of the experiments that exhibited null results. As well, for those studies replicated, the effect sizes were smaller than initially reported.

In 2015, *Nature* conducted a survey of over 1500 researchers on issues related to reproducibility. In the fields of biology and medicine, over 50% of researchers failed to replicate their own experiments, and at least 60% reported failing to reproduce the work of someone else. Two-thirds of those surveyed also reported establishing procedures in the laboratory to support reproducible work [7]. While some of these numbers seem quite high, this report may also highlight an aspect of the very nature of the scientific method, in which correction within research subfields is a necessary component of validating essential results.

Scientific discourse concerning research results is a natural component of the scientific method. A recent analysis of "disagreement" within four million scientific research articles found that 0.41% of papers published in the broad category of "biomedical and health sciences" references disagreements with prior published work [8]. This disagreement with prior literature was categorized as either "paper-level disagreement" or "community-level disagreement" and included a definition of disagreement that encompassed discussion of controversy, dissonance, explicit disagreement with prior work, or lack of consensus with prior work or works [8]. These and other data naturally lead to a discussion of whether this is acceptable "noise" within the scientific community or not. Hypotheses and theorems that may be supported by evidence can always be toppled by new, stronger data or ideas. Providing new evidence that questions prior ideas is an imperative role the research community plays in monitoring its own advancement.

Alternative approaches have been taken to address the reported reproducibility crisis. Retraction Watch began as a citizen science website in 2010 to document and track retractions of research papers or other scholarly work in research. Between the beginning of 2012 and the end of September 2022, over 1200 research articles related to the keyword "genetics" had been retracted due to concerns or errors with the data. Similar results occur when searching the same time period for papers related to "cancer" or "oncology". Dr. Elisabeth Bik has made a second career out of identifying fraudulent research via her Science Integrity Digest, highlighting manipulated figure images on her social media accounts [9]. In 2019, she led a study examining 960 research papers published in *Molecular and Cellular Biology* between 2009 and 2016 and found that 6% had inappropriately duplicated figure images [10,p. 20]. This was a follow-up to an earlier study of over 20,000 papers published within 40 journals between 1995 and 2014. She and her colleagues found that almost 4% of these papers had problems with one or more figures and that at least half of these, 2% of all the papers, had evidence of visual manipulation [11].

In the field of genetics and genomics, structural problems contribute to a lack of rigorous research practice. Historically, nearly 96% of all participants in all genome-wide association studies (GWAS) are of European ancestry, with a paltry 3% of Asian ancestry being the next most represented ancestral population [12]. Lack of ancestral representation in GWAS and related genomic analyses limits the ability to identify physiologically- or disease-relevant variation in the human genome—the true variation the human genome is not being accurately captured. How can geneticists infer the genetic contributors to disease processes if the complexity of variation that contributes to the said diseases is largely ignored? Presently, the shocking lack of representation in data sets limits the ability to extrapolate our understanding of genetic contributors to disease to populations outside of Western European ancestries.

Initiatives such as the National Institutes of Health's (NIH) *All of Us* research program has been developed to increase the diversity of biomedical research studies [13] and promote new opportunities

to expand our knowledge about genomic diversity. The H3Africa (Human Hereditary and Health in Africa) Initiative is a leading consortium of researchers and laboratories in Africa to further address the disparity in our knowledge about variation in the human genome [14]. While these essential databases and others like it catch up on the collection of diverse biospecimens, detailed health history, and necessary representative sample sizes, geneticists have based most of the field's knowledge of fundamental diseases processes on the Western European genome.

Numerous statistical approaches exist for inferring associative and causal DNA variants related to disease development, environmental response, and other physiological pathways. These approaches include polygenic risk scoring (PRS), Mendelian randomization, estimates of heritability, genome-wide copy number variant (CNV) analysis, identifying variation in allele variation to estimate human migration, and others [15,16]; however, the past decade has seen GWAS dominate this realm of "big data" statistical genomic research. Most of these analyses are built on the foundation of databases such as the UK Biobank, which have >90% European ancestry in their sampled populations [17]. The 1000 Genomes Project Consortium is more diverse, with samples from 26 different ethnic populations; however, there are on average only ~100 samples per population in the database [18–20]. The small sample size per ethnic population means that most studies will be severely underpowered, limiting the ability to detect novel variants and smaller, but still physiologically relevant, effect sizes.

There are thus a number of additional factors contributing to the rigor and reproducibility crisis in biomedical research, with specific concerns for genetic and genomic researchers. These factors include funding challenges and an unhealthy culture around publishing results, structural challenges in genetic research and diverse sample collection/patient recruitment (and ethical compensation), and a lack of rigorous reporting and data sharing standards. These factors and more are detailed in "What are the contributing factors to the reproducibility crisis?" section and discussed at length.

This textbook endeavors to identify and address technical and methodological issues in genomic research that negatively impact reproducibility of data, and rigorous research practices. Additionally, corollary factors that impact rigor and reproducibility in research are discussed, including: improving genetic education at the secondary and post-secondary levels as well as in graduate training; communication in collaboration and study design; methodology and data sharing; and general transparency and open science practices. These considerations together strengthen the methodology of a research study and increase the confidence and replicability of results [2].

The issue of waning public trust in scientific research

Unfortunately, the era of social media and sensationalized headlines, combined with financial interests by competing groups, including "Big Natural" (a term coined by Dr. Jen Gunter, a self-proclaimed fighter for evidence-based women's health), leads to disagreement within and beyond the scientific community. The scientific process is naturally self-correcting. As evidence accumulates and results are replicated (or not), every bit of incorrect, non-rigorously conducted or reported research that makes its way to the public prior to being identified as such contributes to the confusion and misinformation campaigns that fuel the media's economic engine (including social media influencers), sowing distrust among the public. The time and space to conduct science and verify results has thus shrunk considerably and demands that researchers adhere to the highest standards of rigorous research and reporting (see case studies in Box 1.1 and Box 1.2 for more).

Box 1.1 The SARS-CoV-2 Pandemic

The Coronavirus Disease 2019 (COVID-19) pandemic put the scientific method under immense public scrutiny, changing perceptions globally of what can be accomplished when researchers are provided adequate funding resources, minimal bureaucratic hurdles, and practice Open Science. Unfortunately, the push to publish COVID-19 related information also meant that a small percentage of these published papers (72 papers, or 0.03%, at the time of this writing) were later found to be inaccurate [21]; two of these retractions came from high-profile peer-reviewed journals (*The Lancet* and *New England Journal of Medicine*). For members of the public who understand this to be part of the self-monitoring and self-correcting aspect of the scientific method, changing information based on new data strengthens their belief in the biomedical research machine. In contrast, for those who already feel alienated or lack familiarity with the scientific method or the wider biomedical establishment, changing discourse can breed discomfort and fear. The ongoing societal discourse between researchers promoting their work, the non-scientific public, and advancement of misinformation campaigns has both helped and hindered the global understanding of the scientific method at large, and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

The genomic sequence of SARS-CoV-2 had been identified and published by the end of January 2020, just months into the early stages of the pandemic [22,23], paving the way for a deeper understanding of the nature of the virus and the development and testing of multiple COVID-19 vaccines a few months later. Within 10 months of the first publicly-confirmed case of COVID-19, there were over 125,000 scientific articles published in the scientific literature, of which 30,000 were on preprint services such as the bioRxiv and medRxiv [24]. It was an incredible burst of scientific focus, discovery, and examination.

The public understanding of COVID-19 early in the pandemic was shaped by these preprint servers. Social media and news reporting of preprint COVID-19 findings escalated quickly during the spring of 2020 [25,26] and public understanding and misinformation was influenced by where the public accessed COVID-19-related information [27]. Additionally, journalistic reporting and public misunderstanding about the differences between "preprint" manuscripts and "peer-reviewed" articles fueled misinformation about both COVID-19 itself, and the scientific need to use preprints for rapid sharing of new results, while still waiting for the formal peer review process to be conducted [28,29].

For example, early preprint manuscripts in bioRxiv suggested that the COVID-19 spike protein had genetic sequence similarities with several human immunodeficiency virus (HIV) proteins, which were unlikely to have evolved naturally, suggesting that SARS-CoV-2 might have been engineered [30]. The paper was quickly retracted given the numerous issues with the sequencing approach, the data produced, and its analysis. Nonetheless, conspiracy theorists, and individuals who stood to gain financially from dissemination of misinformation/conspiracies, continued to use preprint articles like this one to promote COVID-19 misinformation and generate public distrust around COVID-19 research, and the medical establishment at large.

This highlights a delicate balance between public engagement with open-source, preprint scientific research and the time it takes for researchers to validate, correct, and review new scientific literature. Further discussion has been called for regarding use of the term "preprint" in news reports on PDFs uploaded to preprint servers so that it is clearer to the non-scientific community that peer review and validation of the results are still required [25]. Mainstream news media seems to be generally cognizant of this important difference and journalists are improving with their adherence to highlight that a preprint article is a non-peer-reviewed PDF published online. Given the accessibility to and rapid promotion of preprint manuscripts, peer-reviewed validation of research within the genetics and genomics community will ultimately have to catch up to insulate against misinformation.

Within the genetics research community, safeguards have been used to validate sequences from SARS-CoV-2 samples and must continue to be used efficiently. The National Center for Biotechnology Information (NCBI) began using the Viral Annotation DefineR (VADR) system to analyze SARS-CoV-2 systems to ensure sequence quality [31]. As well, the NIH hosts an open-access data dashboard to support COVID-19 researchers, including access to the COVID-19 Genome Sequence Dataset to submit sequencing information to the Short Read Archive hosted by NCBI, or the GISAID database supported by Freunde von GISAID e.V. and other partners. These repositories are instrumental in helping the scientific community validate sequencing findings and results, identify novel SARS-CoV-2 variants, as well outline what must be identified in related preprint manuscripts so as to inform journalists and others reporting the results of a particular study.

Box 1.2 The Advancement of CRISPR

Aside from polymerase chain reaction (PCR), nothing has ushered in a tsunami of new genomic and molecular biology research more than the development of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) gene editing [32]. CRISPR has already revolutionized approaches to therapy in the clinic to treat sickle cell anemia and β-Thalasemmia [33] and cancer [34], establish new crops [35], and pave our way of understanding our own development [36,37]. There have also been significant advances in the methodologies of using CRISPR in the laboratory and clinic, including the expansion into several different types of CRISPR-associated (Cas) proteins, the ability to make precise single-base edits, and editing of RNA transcripts [38].

Given the fundamental nature and power of CRISPR approaches, there has been considerable debate within the scientific and public communities on how best to use these potentially generation-altering genomic tools. There are no definitive answers on how best to juggle the moral and ethical implications of CRISPR alterations with the goal of improving human and agricultural health and well-being. This is further complicated by using CRISPR to study embryonic development [39] or editing the human germline.

In 2018, researcher He Jiankui announced the birth of the first human babies born with germline genetic modifications using CRISPR [40]. This news sent the world into shock given that the procedures for the use of CRISPR for heritable transmission in humans had hardly been formalized, or even agreed upon in the international community. The three babies born in China exemplifies one of the major ethical debates in the public related to genomic research. What began just over a decade ago in bacterium has now influenced the lives of children born without a say in the procedure performed on them. Regardless of where geneticists fall on the spectrum of the acceptable use of CRISPR gene editing, answers must be found on a number of issues, including:

- Who should govern the use of CRISPR in the research environment?
- Who has a say in what types of cells are used and what types of experiments are performed using CRISPR?
- How do we navigate off-target DNA edits and management of resources to validate new approaches? [41,42]
- What roles do researchers and the media play in communicating novel findings?
- What role should CRISPR-regulated gene drives play in shaping or modifying the environment? [43]

As more and more research becomes accessible, more and more the non-scientific public will need to be educated on this issue to ensure productive public discourse when answering these and other questions.

What are the contributing factors to the reproducibility crisis?

The scientific method naturally leads scientists to engage in criticism of one another's work—ideally constructively, although this is not always the case. This self-monitoring dynamic is intended to strengthen our foundational knowledge and percolate interest in new research avenues. Despite informal feedback from colleagues and the formalized peer review process, hundreds of peer-reviewed publications, many on topics related to genetics and genomics, are retracted each year. How is it that so many studies "miss the mark," whether intentionally or unintentionally? Why are inaccuracies or flat-out falsehoods missed? What are the intrinsic factors that influence replication and reproduction of research data—particularly positive experimental results?

From the inception of a research question through to publication and subsequent critiquing by members of the field, each tier along the way contributes to whether a study produces the highest quality research in the most reproducible manner or not. The first of these tiers is to ask the correct question—leading or biased questions will inherently give rise to biased conclusions. Next, it is necessary to conduct a thorough review of the literature, to know what has been done and found previously, where the gaps in our knowledge may exist, and whether any of these previous studies have drawn

conclusions that incongruous with their results, or in the context of the field. Methodology and study design are critical—selecting the appropriate samples, controls, techniques, and tests will strengthen the quality of the data produced and the conclusions that can be drawn.

Data analysis is the next significant point where many research studies stumble, establishing a significant finding where one might not exist due to the use of inappropriate statistical tests. Interpretations of these analyses can be challenging, and at times over-interpretation despite weak evidence leads authors to propose causality where it does not clearly exist. Accurate and transparent reporting of methods and all results (not just the positive results), free sharing of code and data sets used in computational work, and publication of raw (unprocessed) research data (whether through a publisher's data repository, supplementary results, GitHub, or via a privately hosted website) is a basic tenet of rigorous, open science. Finally, we come to peer review and publication—where, in theory, oversights or flat-out mistakes in the aforementioned stages should be caught, revised, and re-submitted for review. Unfortunately, given the complexity of much genetic and genomic research, and the time pressures faced by researchers, peer review is not the *silver bullet* to solving the rigor and reproducibility crisis. The most crucial of these stages and factors affecting reproducibility are expanded upon below.

Numerous methodological factors contribute to the validity of a research study. Munafo et al. reviews that factors such as publication bias, failure to control for bias, low statistical power, poor quality control, and *P*-hacking can all contribute to undermining the validity of research studies and inhibiting other laboratories' ability to reproduce work [2]. It is also becoming increasingly important for geneticists to have at least some foundational understanding of biostatistics and statistical science. Appropriate tests must be chosen, given a specific context, for correction of false positives [44], variant imputation [45,46], population structure and confounding variables [47], or even within pipelines to account and control for internal technical errors caused by the sequencing platform [48]. There can also be important considerations when combining different data sets and admixture of samples [49], or even deciding upon an appropriate threshold for significance [50].

As mentioned earlier in this chapter, the lack of diverse representation in most GWAS and/or study populations can also impair efforts to replicate findings. Homogenous cohorts fail to capture functional variants in the human genome that are important for physiological processes or disease progression. Downstream, this homogeneity creates problems when building new protocols or platform technologies for sequencing and variant calling of new samples, as it utilizes assumptions or known variants identified only in a single population. This is especially relevant when using polygenic risk scores (PRS) to assess and predict predisposition to different conditions (discussed further in Chapter 5). Given a majority of PRS calculations were performed using underlying variant data from individuals of European ancestry, PRS in individuals from other backgrounds are less accurate and useful in the clinic [51–53].

A corollary contributing factor to the reproducibility crisis, supplementary to the lab bench itself, is the culture of career advancement within academic research, highlighted by the proverbial "publish or perish" narrative. This narrative and reality in academic science pressures early career investigators to show their research productivity by publishing multiple papers as a means of establishing job security. While there are many other components to the tenure package in academia, the ability to show productivity from grant funding and the ability to deliver research results is the primary consideration for tenure review committees. While it seems superficially sensible that promotion should be tied to scholarship, particularly the ability to conduct and publish impacting research, there is a disconnect between this requirement for job security and the culture of how research is reported in the literature.

The primary example of this bias in published literature is the fact that there exists a systemic reporting bias that emphasizes positive results in peer-reviewed literature and disfavors the reporting of negative results, even among biomedical and clinical research trials [54,55]. In turn, this influences the approaches that investigators (particularly early-career investigators) take to validate their research, knowing their livelihoods and those in their labs are dependent upon showing successful outcomes in their work. This disconnect can be perpetuated by review, promotion, and tenure (RPT) committees dependent on the institutional metrics used to define the scholarly success of faculty members under consideration for promotion.

Inappropriate measures of scholarship, such as impact factor (IF) or rewarding quantity over quality (which can lead to a lack of reproducibility) can also inappropriately incentivize biomedical researchers to publish work that reinforces job protection and less-than-excellent scholarship [56–58]. Responsibility for training the next generation of researchers also falls heavily on principal investigators. Genetic researchers at all levels, and particularly research associates and principal investigators, can help develop strong scholarly habits in trainees via the demonstration and reinforcement of responsible, rigorous research conduct. One should encourage open and honest communication regarding reporting preliminary findings and during meetings with collaborators. Further, setting and upholding laboratory policies for recording thorough and accurate lab notes, and reporting research misconduct when it occurs, provide valuable tools and lessons to graduate trainees. The latter requires mandatory and extensive training regarding responsible conduct of research and also requires that trainees are provided institutional and field-specific resources to access when needed [59].

There should also be articulated institutional-specific policies for early-career investigators to follow when questions related to research integrity arise that can be professionally explored without necessarily being automatically punitive. These internal review policies of institutions may also play a role in the repercussions for researchers who falsify or fabricate data.

Across US and global institutions, the policies for investigating cases of fabricated or falsified data vary widely. Best practices for reviewing these cases that are more widely adopted may help reduce the frequency of retractions in the scientific literature [60,61]. An analysis of 1316 papers published from US institutions across multiple scientific disciplines found that the competitive environment of the authors' institution biased against reporting negative research results [62]. This and other work has spurred discussion on how best to remedy the bias that influences reliable result reporting.

Some journals have taken a new approach to emphasize the methodology of the science as opposed to the results or findings. *Cell Press*, a peer-reviewed journal within the Elsevier portfolio, launched *STAR Protocols* in 2016 to identify reproducible protocols in the life sciences that were accessible and validated [63]. STAR stands for Structured Transparent Accessible Reproducible, and the journal articles are reviewed by core facility and technologically experienced research scientists. The Center for Open Science initiated the use of **Registered Reports** to re-emphasize peer review on the methodology of the study as opposed to the final results of the analysis.

In a Registered Report, researchers submit their idea and study design for an initial round of peer review, in which reviewers weigh the integrity and strength of the research idea and methodology. If the report passes this round, the paper is conditionally accepted, regardless of the results of the study, pending adherence to the reviewed protocol [64,65]. Select journals will accept and publish genetic studies that are pre-registered reports as part of their publishing model, include *Scientific Reports*, *PLOS ONE, PLOS Biology*, *BMC Biology*, and *BMC Medicine*.

eLife recently adopted a new peer review protocol that requires all reviewed articles to first be published as a preprint. Next, the reviewed article is automatically published by the journal regardless of the peer review process. This new form of acceptance also includes the views of the reviewing experts, those who have discussed the work on the preprint forums, and the author's reply (if necessary). This radical change has removed the accept/revise/reject model of formal peer review [66] and already sparked considerable and healthy debate within the scientific community.

Given the complex nature of some genomic analysis, additional resources will be needed to help trainees and early-career investigators develop the necessary intuition and skillset to ask appropriate questions that challenge the integrity of a given methodology, whether with their own work or another's. These questions should become second nature for newly trained researchers; perhaps as ingrained into graduate training as is the emphasis on identifying a research question, developing a testable hypothesis, or designing and analyzing a more inclusive (diverse) cohort. If there is more openness up front on how to develop the best methodological approach to a particular experiment or question, or how to best review it, there will be fewer concerns about the results if they are not able to be replicated elsewhere.

The societal importance of open science

> When Jonas Salk was asked who owned the patent to his new polio vaccine, he famously replied, "Well, the people, I would say. There is no patent. Could you patent the sun?"

In all the years since 1955, Dr. Jonas Salk's idea that his and his team's science be available solely for the betterment of humanity is still a high bar to achieve given the current systemic infrastructure of research, publishing, patenting, and health care. With the advent of modern technologies that reduce cost and time, the ideal of "open science" has inspired the creation of large, public, and free databases that have promoted research and considerable secondary research worldwide.

Unfortunately, given the enormous influence of profit-driven privatization of medical care and insurance, particularly in the United States, and elsewhere in the world, there are many economic factors that prevent the latest breakthroughs from establishing themselves for free or with widespread usage in the public domain. One need to look no further than the patent disputes between MIT's Broad Institute and the University of California, Berkeley (alongside Dr. Emmanuelle Charpentier) regarding ownership of CRISPR gene-editing technology—a legal drama that continues to unfold. Each institute is keenly aware of the economic boon from owning control of CRISPR and the downstream licensing of this approach, and this is just within the United States. The issue becomes even more complex when looking at patent ownership of CRISPR technologies in the European Union and elsewhere.

Dramatic steps have been taken toward the democratization of science and unrestricted access of research results and large data sets. A prime example of this is the UK Biobank, an open-access database with greater than half a million genomes (with phenotypic data), to which any qualified scientist on the planet can apply for ethical approval access. The UK Biobank is a not-for-profit organization, supported by various levels of UK government and charitable foundations. Although not a perfect resource—the database lacks samples of ethnic diversity (as discussed above)—it continues to add new genomes regularly and provided a wealth of information to mine for large-scale genomic studies. An unusual example of the democratization of biology comes in the form of 3D printing technologies, which are increasingly allowing researchers to design tools or modify those that they have already,

eliminating the high costs of biotech sales and increasing specificity tailored to their needs. In addressing public access to published-behind-a-paywall articles, all research that is federally-funded by the United States government will be required to be immediately available and open access upon publication by 2026 [67]. Steps like these ensure that all researchers, as well as the general public, have access to essential data and analysis as quickly as possible.

The field of genomic research has seen an exponential growth in the amount of data generated and made available to researchers and the public. Open science and data sharing agreements have become increasingly important in managing this data. One of the key challenges is balancing the need for data sharing with protecting patient privacy. The 1000 Genomes Project Consortium [18] and the National Cancer Institute's Genomic Data Commons [68] are two examples of successful data sharing initiatives.

The Genomic Data Commons integrates clinical data from individual studies by harmonizing inputs on sample collection, the alignment of sequencing data to a common reference genome, and standardizing protocols on variant calling, and other metrics. There are also controlled and restricted data sets within this public database (and others) that are curated in accordance with the informed consent documents or other guidelines delineated when participants are recruited into participating studies. This identifiable data may be embargoed or behind a secure wall such that only those who apply to access this data are granted permission to use it. While not entirely open access, these restrictions reflect necessary precautions needed for patient privacy.

Data availability is also determined by the country hosting the database. In the United States, there are numerous federal and state laws that regulate the collection, usage, and disclosure of genomic data. For example, The Genetic Information Nondiscrimination Act (GINA) prohibits employers, health insurance companies, and others from using genetic information to discriminate against individuals. The European Union has adopted the General Data Protection Regulation (GDPR) which protects the privacy of personal data, including genomic data [69]. The GDPR requires that individuals must provide informed consent for the collection and use of their data, and it gives individuals the right to access, rectify, and erase their data. The GDPR also requires that organizations implement appropriate technical and organizational measures to protect personal data. The law prohibits processing this data in such a way that could even indirectly reveal sensitive information about an individual.

In China, the Cybersecurity Law, Data Security Law, and Personal Information Protection Law (PIPL) have been implemented to govern how personal identifiable information (both biological and digital) are collected, protected, and stored in China. These regulations also delineate that consent for this information to be collected must be freely given and informed and that it can be withdrawn.

While these laws have made it challenging for geneticists and researchers to access and use genetic data [70], they are essential to protect personal information in a rapidly changing research environment. Additional guidance has been needed for open access of genetic data beyond these laws. For example, in the United States, there has been historically many cases of data mismanagement and lack of consent when it comes to the collection and use of samples from indigenous communities, and other racial and ethnic populations historically underrepresented in genomic studies. New guidelines that focus on trust, accountability, and equity must be implemented to ensure protection of this information and safeguard against sample misuse, along with including the input of the participants in the study who are providing the samples [71]. Data consortiums must also be sensitive to our changing understanding of the intersection of race, ethnicity, and ancestry, especially when samples are being collated together from different genomic databases [72,73].

These and other guidelines should always be continually revisited to ensure equitable access and protection of genomic information. Ideally, open science ensures that researchers and bioethicists always have the opportunity to shore up problems in research pipelines, the process of study participant recruitment, consent, and engagement, and in reporting analysis outcomes.

The non-scientific public must also continue to have a stronger voice in how this data is used and discussed. Social media platforms such as Twitter, Facebook, and Mastodon allow researchers to engage directly with the public and the media. In the first months of the COVID-19 pandemic, hundreds of thousands of tweets on Twitter discussed a variety of topics related to the information from and perception of the Centers for Disease Control and Prevention (CDC) regarding COVID-19. The most discussed topics included the credibility of the CDC and the CDC guidelines related to COVID-19 exposure and response [74].

This rapid fire promotion of the latest in scientific discovery is a boost to equitable access to research results and informed policy but can also promote mistrust in the process of science and aid in the spread of misinformation or false information [75,76]. Twitter bots and other malware can spread misinformation or sow the appearance of disagreements within a scientific field when there is large consensus, as what has happened concerning the discussion focused on the safety and efficacy of vaccinations [77].

Genetic and genomic studies are not immune to these trends. When news of He Jinkaiu's experiment using CRISPR and the birth of the first CRISPR-edited humans, Twitter, Chinese social media platform Weibo, and other social media platforms explored with discourse related to the ethical controversy and societal implications of its use [78,79] (see Box 1.2). These conversations appear to be linked with the news cycle in that conversations can be tied with when news breaks related to a specific event or key development in genetics research [78,80].

Additional consequences of genetic and genomic information being so easily accessible have extended far beyond the halls of academia and industry. Direct-to-consumer (DTC) DNA testing has grown in the last decade and contributed to mainstream discussion of genetic variation, ancestry, and susceptibility to disease. However, not all of the perceived health information related to some of these products are discussed by trained professionals, which opens the public discourse up for the spread of misinformation or basing healthcare decisions based on non-clinical test results [81–83].

Participants of DTC DNA testing are also concerned about opaque privacy protection related to their DNA testing results [84]. DTC testing has influenced family dynamics and relationships when ancestry results return, often without much support from the company providing the service [85]. There are also questions concerning who can give permission to have their DNA tested. This is a particularly complex issue when that individual does not know or authorize the test or is deceased [86]. The results of these analyses can have profound consequences and the impacts on society are still not completely understood.

Arguably one of the most controversial cases of DNA privacy in DTC testing is the use of genetic test results by law enforcement. In 2018, news broke in the United States that the famous Golden State Killer, a serial killer who committed murders in the 1970s, had been identified by police by using the public genealogy website GEDMatch [87]. Law enforcement officials had uploaded DNA from a crime scene and identified a relative of the killer in GEDMatch, ultimately arresting a retired police officer who had committed those terrible crimes. As an additional consequence, the case immediately brought up questions related to the ethical use of DTC testing, including data privacy, public safety, DNA ownership, and other complicated bioethical questions. These questions are further confounded when

weighing personal privacy and protection versus public safety, including ensuring criminals are found. Since 2018, GEDMatch and other genealogy databases have helped solve hundreds of cold cases and crimes.

Negotiating these and other complicated bioethics of genetic research is not formally part of the training of many geneticists in the US and around the world. The NIH has mandated that institutions receiving NIH funding implement RCR training for grant awardees and trainees [88]. RCR training can highlight many different issues including navigating trainee power dynamics, responsible data collection and reporting, conflict of interest, the peer review process, and even "the scientist as a responsible member of society, contemporary ethical issues in biomedical research, and the environmental and societal impacts of scientific research" [88].

However, institutions are generally free to implement RCR training as they see fit, and there is little uniformity across the US or within the international community. There should be incentives at the institutional or national level in graduate training and with early-career faculty development that stresses the importance of a societal-conscious biomedical researcher. Given genetic technology and discovery has become a part of everyday conversation, additional training is needed to help researchers navigate how to discuss their work with a broad and diverse community. Bioethics, rigorous methods, and RCR need to become more integrated into undergraduate and graduate training such that researchers are prepared for these conversations either among themselves, with lawmakers or other members of society, or even within their or social networks of family and friends.

Conclusions

Given the rapid pace of genetic research, it is likely that exciting new advancements in our understanding of the genome will continue to emerge, along with bold interventions in clinical practice. These developments may have unforeseen ramifications, making it critical for geneticists, clinicians, trainees at all levels, patients, and the public to have a voice in how we apply and expand our knowledge. The emerging use of artificial intelligence, like ChatGPT and other AI-driven programs, are rapidly gaining traction in numerous software platforms. These AI programs are in their infancy—the "training" stage—but this is a critical time for AI as the data sets used for training will inform the biases inherent to these platforms. The implications are enormous and wide-reaching in all fields in the context of scientific writing. For example, the ability for an AI to produce scientific literature that *sounds* correct but in fact misconstrues the facts or simply is incorrect leads to an enormous black box about regulating the use of AI in preparation of manuscripts and other publications. Just as this book was preparing to go to press, ChatGPT and other AIs took the internet by storm, so much so that Italy temporarily banned ChatGPT [89] and publishers were forced to quickly respond with guidance to authors on the matter. Elsevier Group (the publisher of this book) issued guidelines in March 2023 stating:

> Where authors use generative AI and AI-assisted technologies in the writing process, these technologies should only be used to improve readability and language of the work….Authors should disclose in their manuscript the use of AI and AI-assisted technologies and a statement will appear in the published work. [66].

In spirit of these guidelines, the authors of this manuscript can reveal that the fourth paragraph of the prior section in this chapter ("The field of genomic research has seen an exponential growth….)" was

imported into ChatGPT to improve readability, and as an example of the power of AI to write scientific works that are indistinguishable from human-authored work.

To better facilitate public discourse of genetic research, it is imperative that the scientific literature reflect the highest level of rigorous methodology. As evidenced by the daily updates of our knowledge of SARS-CoV-2 during the COVID-19 pandemic (see Box 1.1) and the growing clinical use of CRISPR gene-editing (Box 1.2), researchers must ensure that the information they bring forth is meritorious and reproducible, with a responsibility to both scientific and broader communities. The era of open science offers a unique opportunity for collaboration and encourages researchers to work together to define best practices in order to improve the transparency and accessibility of research outcomes.

Reproducible research does not necessarily mean that the results of any given experiment or project will always be correct. Rather, it endeavors to foster the careful consideration required such that the underlying hypothesis, approach of testing the said hypothesis, and the data collected and analyzed are meaningfully interpretable. Geneticists and researchers should approach their work such that it can grow with the changing knowledge of the community at large and that others can go back to ensure our bedrock principles and knowledge are sturdy.

The scientific research enterprise is flawed in that it is limited in part by our preceding knowledge of the world, and in part by the naïve mistakes of the untrained or ignorance of those willing to take short cuts. The convalescence of these aspects can lead to incorrect scientific conclusions, which are at times inappropriately disseminated via the use of preprint servers and AI-supported technologies before researchers are able to discuss and self-correct the science. While there are many ways to tackle these issues to ensure progress in our work and for the betterment of the society, they can be summarized into three strategic goals that the genetic research community should always strive for:

1. The genetics research community should always work to improve the general public's understanding of the scientific process so that open science and public discourse are less reactionary or misinformed.
2. The genetics research community should continue to establish reproducible research practices to strengthen the research findings and make them more representative of the diverse global population.
3. The genetics research community should promote the development of strong science communication skills within the next generation of the research and clinical workforce.

This chapter has outlined a few of the individual and collective actions that can be taken to achieve these aforementioned goals. Institutional and departmental commitment to these or similar ideals will also solidify the genetic research infrastructure as a whole and reinforce the need to continue to execute strong research practices. The subsequent chapters in this textbook are meant to provide a deeper knowledge into reproducible research practices using a variety of widely used approaches in genetics and genomics, from PCR to CRISPR. Additionally, this textbook also provides guidance on how faculty, mentors, or others in instructional positions can infuse and promote rigorous practices into their work and curriculums so that future research trainees achieve the highest standard of reproducible research.

By instilling these practices at all levels of the scientific enterprise, we can continue to push our knowledge of genetics in new and meaningful directions, helping researchers achieve the goal of their studies being peer reviewed, published, and cited!

References

[1] J.P.A. Ioannidis, D. Fanelli, D.D. Dunne, S.N. Goodman, Meta-research: evaluation and improvement of research methods and practices, PLoS Biol. 13 (2015) e1002264, https://doi.org/10.1371/journal.pbio.1002264.

[2] M.R. Munafò, B.A. Nosek, D.V.M. Bishop, K.S. Button, C.D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J.J. Ware, J.P.A. Ioannidis, A manifesto for reproducible science, Nat. Hum. Behav. 1 (2017) 1–9, https://doi.org/10.1038/s41562-016-0021.

[3] H. Shen, Meet this super-spotter of duplicated images in science papers, Nature 581 (2020) 132–136, https://doi.org/10.1038/d41586-020-01363-z.

[4] National Academies of Sciences, Engineering, and Medicine, Understanding reproducibility and replicability, in: Reproducibility and Replicability in Science, The National Academies Press, Washington, DC, 2019, https://doi.org/10.17226/25303.

[5] J.P.A. Ioannidis, Why most published research findings are false, PLoS Med. 2 (2005) e124, https://doi.org/10.1371/journal.pmed.0020124.

[6] Science, C. for O, Reproducibility Project: Cancer Biology, 2023 (WWW Document) https://www.cos.io/rpcb. (Accessed 27 February 2023).

[7] M. Baker, 1,500 scientists lift the lid on reproducibility, Nature 533 (2016) 452–454, https://doi.org/10.1038/533452a.

[8] W.S. Lamers, K. Boyack, V. Larivière, C.R. Sugimoto, N.J. van Eck, L. Waltman, D. Murray, Investigating disagreement in the scientific literature, elife 10 (2021) e72737, https://doi.org/10.7554/eLife.72737.

[9] Science Integrity Digest, Science Integrity Digest, Sci. Integr. Dig, 2022 (WWW Document) https://scienceintegritydigest.com/. (Accessed 27 February 2023).

[10] E.M. Bik, F.C. Fang, A.L. Kullas, R.J. Davis, A. Casadevall, Analysis and correction of inappropriate image duplication: the molecular and cellular biology experience, Mol. Cell. Biol. 38 (2018) e00309–e00318, https://doi.org/10.1128/MCB.00309-18.

[11] E.M. Bik, A. Casadevall, F.C. Fang, The prevalence of inappropriate image duplication in biomedical research publications, MBio 7 (2016) e00809–e00816, https://doi.org/10.1128/mBio.00809-16.

[12] M.C. Mills, C. Rahal, The GWAS diversity monitor tracks diversity by disease in real time, Nat. Genet. 52 (2020) 242–243, https://doi.org/10.1038/s41588-020-0580-y.

[13] All of Us Research Program Overview, Us Res. Program NIH, 2020 (WWW Document) https://allofus.nih.gov/about/program-overview. (Accessed 27 February 2023).

[14] About – H3Africa, 2023. URL https://h3africa.org/index.php/about/. (Accessed 27 February 2023).

[15] N. Chatterjee, J. Shi, M. García-Closas, Developing and evaluating polygenic risk prediction models for stratified disease prevention, Nat. Rev. Genet. 17 (2016) 392–406, https://doi.org/10.1038/nrg.2016.27.

[16] P.M. Visscher, N.R. Wray, Q. Zhang, P. Sklar, M.I. McCarthy, M.A. Brown, J. Yang, 10 years of GWAS discovery: biology, function, and translation, Am. J. Hum. Genet. 101 (2017) 5–22, https://doi.org/10.1016/j.ajhg.2017.06.005.

[17] T.A. Manolio, Using the data we have: improving diversity in genomic research, Am. J. Hum. Genet. 105 (2019) 233–236, https://doi.org/10.1016/j.ajhg.2019.07.008.

[18] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (2015) 68–74, https://doi.org/10.1038/nature15393.

[19] S. Fairley, E. Lowy-Gallego, E. Perry, P. Flicek, The international genome sample resource (IGSR) collection of open human genomic variation resources, Nucleic Acids Res. 48 (2020) D941–D947, https://doi.org/10.1093/nar/gkz836.

[20] P.H. Sudmant, T. Rausch, E.J. Gardner, R.E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M.H.-Y. Fritz, M.K. Konkel, A. Malhotra, A.M. Stütz, X. Shi, F.P. Casale, J. Chen, F. Hormozdiari,

G. Dayama, K. Chen, M. Malig, M.J.P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H.Y.K. Lam, X.J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J.M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R.A. Gibbs, G. Marth, C.E. Mason, A. Menelaou, D.M. Muzny, B.J. Nelson, A. Noor, N.F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E.E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A.A. Shabalin, A. Untergasser, J.A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M.A. Batzer, S.A. McCarroll, 1000 Genomes Project Consortium, R.E. Mills, M.B. Gerstein, A. Bashir, O. Stegle, S.E. Devine, C. Lee, E.E. Eichler, J.O. Korbel, An integrated map of structural variation in 2,504 human genomes, Nature 526 (2015) 75–81, https://doi.org/10.1038/nature15394.

[21] Retracted Coronavirus (COVID-19) Papers, 2023. https://retractionwatch.com/retracted-coronavirus-covid-19-papers/.

[22] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W.J. Liu, D. Wang, W. Xu, E.C. Holmes, G.F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, Lancet 395 (2020) 565–574, https://doi.org/10.1016/S0140-6736(20)30251-8.

[23] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E.C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China, Nature 579 (2020) 265–269, https://doi.org/10.1038/s41586-020-2008-3.

[24] N. Fraser, L. Brierley, G. Dey, J.K. Polka, M. Pálfy, F. Nanni, J.A. Coates, The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape, PLoS Biol. 19 (2021) e3000959, https://doi.org/10.1371/journal.pbio.3000959.

[25] R. Ravinetto, C. Caillet, M.H. Zaman, J.A. Singh, P.J. Guerin, A. Ahmad, C.E. Durán, A. Jesani, A. Palmero, L. Merson, P.W. Horby, E. Bottieau, T. Hoffmann, P.N. Newton, Preprints in times of COVID19: the time is ripe for agreeing on terminology and good practices, BMC Med. Ethics 22 (2021) 106, https://doi.org/10.1186/s12910-021-00667-7.

[26] S.L. Taneja, M. Passi, S. Bhattacharya, S.A. Schueler, S. Gurram, C. Koh, Social media and research publication activity during early stages of the COVID-19 pandemic: longitudinal trend analysis, J. Med. Internet Res. 23 (2021) e26956, https://doi.org/10.2196/26956.

[27] D. De Coninck, T. Frissen, K. Matthijs, L. d'Haenens, G. Lits, O. Champagne-Poirier, M.-E. Carignan, M.D. David, N. Pignard-Cheynel, S. Salerno, M. Généreux, Beliefs in conspiracy theories and misinformation about COVID-19: comparative perspectives on the role of anxiety, depression and exposure to and Trust in Information Sources, Front. Psychol. 12 (2021) 646394, https://doi.org/10.3389/fpsyg.2021.646394.

[28] L. Brierley, Lessons from the influx of preprints during the early COVID-19 pandemic, Lancet Planet. Health 5 (2021) e115–e117, https://doi.org/10.1016/S2542-5196(21)00011-5.

[29] M.S. Majumder, K.D. Mandl, Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility, Lancet Glob. Health 8 (2020) e627–e630, https://doi.org/10.1016/S2214-109X(20)30113-3.

[30] P. Pradhan, A.K. Pandey, A. Mishra, P. Gupta, P.K. Tripathi, M.B. Menon, J. Gomes, P. Vivekanandan, B. Kundu, Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag, bioRxiv (2020) 927871, https://doi.org/10.1101/2020.01.30.927871.

[31] A.A. Schäffer, E.L. Hatcher, L. Yankie, L. Shonkwiler, J.R. Brister, I. Karsch-Mizrachi, E.P. Nawrocki, VADR: validation and annotation of virus sequence submissions to GenBank, BMC Bioinformatics 21 (2020) 211, https://doi.org/10.1186/s12859-020-3537-3.

[32] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J.A. Doudna, E. Charpentier, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity, Science 337 (2012) 816–821, https://doi.org/10.1126/science.1225829.

[33] H. Frangoul, D. Altshuler, M.D. Cappellini, Y.-S. Chen, J. Domm, B.K. Eustace, J. Foell, J. de la Fuente, S. Grupp, R. Handgretinger, T.W. Ho, A. Kattamis, A. Kernytsky, J. Lekstrom-Himes, A.M. Li, F. Locatelli, M.Y. Mapara, M. de Montalembert, D. Rondelli, A. Sharma, S. Sheth, S. Soni, M.H. Steinberg, D. Wall, A. Yen, S. Corbacioglu, CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia, N. Engl. J. Med. 384 (2021) 252–260, https://doi.org/10.1056/NEJMoa2031054.
[34] A. Dimitri, F. Herbst, J.A. Fraietta, Engineering the next-generation of CAR T-cells with CRISPR-Cas9 gene editing, Mol. Cancer 21 (2022) 78, https://doi.org/10.1186/s12943-022-01559-z.
[35] K. Chen, Y. Wang, R. Zhang, H. Zhang, C. Gao, CRISPR/Cas genome editing and precision plant breeding in agriculture, Annu. Rev. Plant Biol. 70 (2019) 667–697, https://doi.org/10.1146/annurev-arplant-050718-100049.
[36] B. Artegiani, D. Hendriks, J. Beumer, R. Kok, X. Zheng, I. Joore, S. Chuva de Sousa Lopes, J. van Zon, S. Tans, H. Clevers, Fast and efficient generation of knock-in human organoids using homology-independent CRISPR-Cas9 precision genome editing, Nat. Cell Biol. 22 (2020) 321–331, https://doi.org/10.1038/s41556-020-0472-5.
[37] A. Martinez-Silgado, F.A. Yousef Yengej, J. Puschhof, V. Geurts, C. Boot, M.H. Geurts, M.B. Rookmaaker, M.C. Verhaar, J. Beumer, H. Clevers, Differentiation and CRISPR-Cas9-mediated genetic engineering of human intestinal organoids, STAR Protoc. 3 (2022) 101639, https://doi.org/10.1016/j.xpro.2022.101639.
[38] J. Tao, D.E. Bauer, R. Chiarle, Assessing and advancing the safety of CRISPR-Cas tools: from DNA to RNA editing, Nat. Commun. 14 (2023) 212, https://doi.org/10.1038/s41467-023-35886-6.
[39] M. Morrison, S. de Saille, CRISPR in context: towards a socially responsible debate on embryo editing, Palgrave Commun. 5 (2019) 1–9, https://doi.org/10.1057/s41599-019-0319-5.
[40] New Scientist, CRISPR Babies: What's Next for the Gene-Edited Children from Trial in China?, New Scientist, 2022 (WWW Document) https://www.newscientist.com/article/mg25533930-700-whats-next-for-the-gene-edited-children-from-crispr-trial-in-china/. (Accessed 27 February 2023).
[41] P.J. Chen, D.R. Liu, Prime editing for precise and highly versatile genome manipulation, Nat. Rev. Genet. 24 (2023) 161–177, https://doi.org/10.1038/s41576-022-00541-1.
[42] B. Wienert, M.K. Cromer, CRISPR nuclease off-target activity and mitigation strategies, Front. Genome Ed. 4 (2022) 1050507, https://doi.org/10.3389/fgeed.2022.1050507.
[43] W.T. Garrood, N. Kranjc, K. Petri, D.Y. Kim, J.A. Guo, A.M. Hammond, I. Morianou, V. Pattanayak, J.K. Joung, A. Crisanti, A. Simoni, Analysis of off-target effects in CRISPR-based gene drives in the human malaria mosquito, Proc. Natl. Acad. Sci. U. S. A. 118 (2021) e2004838117, https://doi.org/10.1073/pnas.2004838117.
[44] K. Korthauer, P.K. Kimes, C. Duvallet, A. Reyes, A. Subramanian, M. Teng, C. Shukla, E.J. Alm, S.C. Hicks, A practical guide to methods controlling false discoveries in computational biology, Genome Biol. 20 (2019) 118, https://doi.org/10.1186/s13059-019-1716-1.
[45] Q. Sun, W. Liu, J.D. Rosen, L. Huang, R.G. Pace, H. Dang, P.J. Gallins, E.E. Blue, H. Ling, H. Corvol, L.J. Strug, M.J. Bamshad, R.L. Gibson, E.W. Pugh, S.M. Blackman, G.R. Cutting, W.K. O'Neal, Y.-H. Zhou, F.A. Wright, M.R. Knowles, J. Wen, Y. Li, Cystic Fibrosis Genome Project, Leveraging TOPMed imputation server and constructing a cohort-specific imputation reference panel to enhance genotype imputation among cystic fibrosis patients, HGG Adv. 3 (2022) 100090, https://doi.org/10.1016/j.xhgg.2022.100090.
[46] Q. Sun, Y. Yang, J.D. Rosen, M.-Z. Jiang, J. Chen, W. Liu, J. Wen, L.M. Raffield, R.G. Pace, Y.-H. Zhou, F.A. Wright, S.M. Blackman, M.J. Bamshad, R.L. Gibson, G.R. Cutting, M.R. Knowles, D.R. Schrider, C. Fuchsberger, Y. Li, MagicalRsq: machine-learning-based genotype imputation quality calibration, Am. J. Hum. Genet. 109 (2022) 1986–1997, https://doi.org/10.1016/j.ajhg.2022.09.009.
[47] J.H. Sul, L.S. Martin, E. Eskin, Population structure in genetic studies: confounding factors and mixed models, PLoS Genet. 14 (2018) e1007309, https://doi.org/10.1371/journal.pgen.1007309.
[48] Y. Yin, C. Butler, Q. Zhang, Challenges in the application of NGS in the clinical laboratory, Hum. Immunol. 82 (2021) 812–819, https://doi.org/10.1016/j.humimm.2021.03.011.

[49] K.E. Grinde, L.A. Brown, A.P. Reiner, T.A. Thornton, S.R. Browning, Genome-wide significance thresholds for admixture mapping studies, Am. J. Hum. Genet. 104 (2019) 454–465, https://doi.org/10.1016/j.ajhg.2019.01.008.
[50] P.C. Sham, S.M. Purcell, Statistical power and significance testing in large-scale genetic studies, Nat. Rev. Genet. 15 (2014) 335–346, https://doi.org/10.1038/nrg3706.
[51] L. Duncan, H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, B. Domingue, Analysis of polygenic risk score usage and performance in diverse human populations, Nat. Commun. 10 (2019) 3328, https://doi.org/10.1038/s41467-019-11112-0.
[52] A.R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B.M. Neale, M.J. Daly, Clinical use of current polygenic risk scores may exacerbate health disparities, Nat. Genet. 51 (2019) 584–591, https://doi.org/10.1038/s41588-019-0379-x.
[53] Y. Wang, K. Tsuo, M. Kanai, B.M. Neale, A.R. Martin, Challenges and opportunities for developing more generalizable polygenic risk scores, Annu. Rev. Biomed. Data Sci. 5 (2022) 293–320, https://doi.org/10.1146/annurev-biodatasci-111721-074830.
[54] M. Mitra-Majumdar, A.S. Kesselheim, Reporting bias in clinical trials: progress toward transparency and next steps, PLoS Med. 19 (2022) e1003894, https://doi.org/10.1371/journal.pmed.1003894.
[55] E.H. Turner, A. Cipriani, T.A. Furukawa, G. Salanti, Y.A. de Vries, Selective publication of antidepressant trials and its influence on apparent efficacy: updated comparisons and meta-analyses of newer versus older trials, PLoS Med. 19 (2022) e1003886, https://doi.org/10.1371/journal.pmed.1003886.
[56] J.P.A. Ioannidis, S. Greenland, M.A. Hlatky, M.J. Khoury, M.R. Macleod, D. Moher, K.F. Schulz, R. Tibshirani, Increasing value and reducing waste in research design, conduct, and analysis, Lancet 383 (2014) 166–175, https://doi.org/10.1016/S0140-6736(13)62227-8.
[57] E.C. McKiernan, L.A. Schimanski, C. Muñoz Nieves, L. Matthias, M.T. Niles, J.P. Alperin, Use of the journal impact factor in academic review, promotion, and tenure evaluations, elife 8 (2019) e47338, https://doi.org/10.7554/eLife.47338.
[58] D.B. Rice, H. Raffoul, J.P.A. Ioannidis, D. Moher, Academic criteria for promotion and tenure in biomedical sciences faculties: cross sectional analysis of international sample of universities, BMJ 369 (2020) m2081, https://doi.org/10.1136/bmj.m2081.
[59] A.L. Antes, L.B. Maggi, How to conduct responsible research: a guide for graduate students, Curr. Protoc. 1 (2021) e87, https://doi.org/10.1002/cpz1.87.
[60] H. Bauchner, P.B. Fontanarosa, A. Flanagin, J. Thornton, Scientific misconduct and medical journals, JAMA 320 (2018) 1985–1987, https://doi.org/10.1001/jama.2018.14350.
[61] C.K. Gunsalus, A.R. Marcus, I. Oransky, Institutional research misconduct reports need more credibility, JAMA 319 (2018) 1315–1316, https://doi.org/10.1001/jama.2018.0358.
[62] D. Fanelli, Do pressures to publish increase scientists' Bias? An empirical support from US states data, PLoS One 5 (2010) e10271, https://doi.org/10.1371/journal.pone.0010271.
[63] E. Marcus, A STAR is born, Cell 166 (2016) 1059–1060, https://doi.org/10.1016/j.cell.2016.08.021.
[64] C.D. Chambers, E. Feredoes, S.D. Muthukumaraswamy, P.J. Etchells, C.D. Chambers, E. Feredoes, S.D. Muthukumaraswamy, P.J. Etchells, Instead of "playing the game" it is time to change the rules: registered reports at *AIMS Neuroscience* and beyond, AIMS Neurosci. 1 (2014) 4–17, https://doi.org/10.3934/Neuroscience.2014.1.4.
[65] B.A. Nosek, D. Lakens, Registered reports: a method to increase the credibility of published results, Soc. Psychol. 45 (2014) 137, https://doi.org/10.1027/1864-9335/a000192.
[66] Elsevier, Publishing Ethics for Editors, 2023 (WWW Document) https://www.elsevier.com/about/policies/publishing-ethics. (Accessed 21 March 2023).
[67] White House, OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay | OSTP, White House, 2022 (WWW Document) https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/. (Accessed 21 March 2023).

[68] Z. Zhang, K. Hernandez, J. Savage, S. Li, D. Miller, S. Agrawal, F. Ortuno, L.M. Staudt, A. Heath, R.L. Grossman, Uniform genomic data analysis in the NCI genomic data commons, Nat. Commun. 12 (2021) 1226, https://doi.org/10.1038/s41467-021-21254-9.

[69] B.M. Knoppers, A. Bernier, S. Bowers, E. Kirby, Open data in the era of the GDPR: lessons from the human cell atlas, Annu. Rev. Genomics Hum. Genet. 24 (2023) null, https://doi.org/10.1146/annurev-genom-101322-113255.

[70] E.W. Clayton, A.M. Tritell, A.M. Thorogood, Avoiding liability and other legal land mines in the evolving genomics landscape, Annu. Rev. Genomics Hum. Genet. 24 (2023) null, https://doi.org/10.1146/annurev-genom-100722-021725.

[71] M. Hudson, N.A. Garrison, R. Sterling, N.R. Caron, K. Fox, J. Yracheta, J. Anderson, P. Wilcox, L. Arbour, A. Brown, M. Taualii, T. Kukutai, R. Haring, B. Te Aika, G.S. Baynam, P.K. Dearden, D. Chagné, R.S. Malhi, I. Garba, N. Tiffin, D. Bolnick, M. Stott, A.K. Rolleston, L.L. Ballantyne, R. Lovett, D. David-Chavez, A. Martinez, A. Sporle, M. Walter, J. Reading, S.R. Carroll, Rights, interests and expectations: indigenous perspectives on unrestricted access to genomic data, Nat. Rev. Genet. 21 (2020) 377–384, https://doi.org/10.1038/s41576-020-0228-x.

[72] J.K. Wagner, J.-H. Yu, D. Fullwiley, C. Moore, J.F. Wilson, M.J. Bamshad, C.D. Royal, Guidelines for genetic ancestry inference created through roundtable discussions, Hum. Genet. Genomics Adv. 4 (2023), https://doi.org/10.1016/j.xhgg.2023.100178.

[73] A.T. Khan, S.M. Gogarten, C.P. McHugh, A.M. Stilp, T. Sofer, M.L. Bowers, Q. Wong, L.A. Cupples, B. Hidalgo, A.D. Johnson, M.-L.N. McDonald, S.T. McGarvey, M.R.G. Taylor, S.M. Fullerton, M.P. Conomos, S.C. Nelson, Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: experiences from the NHLBI TOPMed program, Cell Genomics 2 (2022) 100155, https://doi.org/10.1016/j.xgen.2022.100155.

[74] J.C. Lyu, G.K. Luli, Understanding the public discussion about the Centers for Disease Control and Prevention during the COVID-19 pandemic using twitter data: text mining analysis study, J. Med. Internet Res. 23 (2021) e25108, https://doi.org/10.2196/25108.

[75] W.-Y.S. Chou, A. Oh, W.M.P. Klein, Addressing health-related misinformation on social media, JAMA 320 (2018) 2417–2418, https://doi.org/10.1001/jama.2018.16865.

[76] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151, https://doi.org/10.1126/science.aap9559.

[77] D.A. Broniatowski, A.M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S.C. Quinn, M. Dredze, Weaponized health communication: twitter bots and Russian trolls amplify the vaccine debate, Am. J. Public Health 108 (2018) 1378–1384, https://doi.org/10.2105/AJPH.2018.304567.

[78] J. Ji, M. Robbins, J.D. Featherstone, C. Calabrese, G.A. Barnett, Comparison of public discussions of gene editing on social media between the United States and China, PLoS One 17 (2022) e0267406, https://doi.org/10.1371/journal.pone.0267406.

[79] C. Ni, Z. Wan, C. Yan, Y. Liu, E.W. Clayton, B. Malin, Z. Yin, The public perception of the #gene EditedBabies event across multiple social media platforms: observational study, J. Med. Internet Res. 24 (2022) e31687, https://doi.org/10.2196/31687.

[80] C.G. Allen, B. Andersen, M.J. Khoury, M.C. Roberts, Current social media conversations about genetics and genomics in health: a twitter-based analysis, Public Health Genomics 21 (2018) 93–99, https://doi.org/10.1159/000494381.

[81] C.H. Basch, G.C. Hillyer, L. Samuel, E. Datuowei, B. Cohn, Direct-to-consumer genetic testing in the news: a descriptive analysis, J. Community Genet. 14 (2023) 63–69, https://doi.org/10.1007/s12687-022-00613-z.

[82] J.S. Roberts, M.C. Gornick, D.A. Carere, W.R. Uhlmann, M.T. Ruffin, R.C. Green, Direct-to-consumer genetic testing: user motivations, decision making, and perceived utility of results, Public Health Genomics 20 (2017) 36–45, https://doi.org/10.1159/000455006.

[83] M. Smith, S. Miller, A principled approach to cross-sector genomic data access, Bioethics 35 (2021) 779–786, https://doi.org/10.1111/bioe.12919.

[84] G.L. Ruhl, J.W. Hazel, E.W. Clayton, B.A. Malin, Public attitudes toward direct to consumer genetic testing, in: AMIA Annu. Symp. Proc. AMIA Symp. 2019, 2019, pp. 774–783.
[85] S. Zhang, When a DNA Test Reveals Your Daughter Is Not Your Biological Child, The Atlantic, 2018 (WWW Document) https://www.theatlantic.com/science/archive/2018/10/dna-test-divorce/571684/. (Accessed 10 March 2023).
[86] S. Zhang, Is DNA Left on Envelopes Fair Game for Testing?, The Atlantic, 2019 (WWW Document) https://www.theatlantic.com/science/archive/2019/03/dna-tests-for-envelopes-have-a-price/583636/. (Accessed 10 March 2023).
[87] S. Zhang, How a Tiny Website Became the Police's Go-To Genealogy Database, The Atlantic, 2018 (WWW Document) https://www.theatlantic.com/science/archive/2018/06/gedmatch-police-genealogy-database/561695/. (Accessed 10 March 2023).
[88] NOT-OD-22-055: FY 2022, Updated Guidance: Requirement for Instruction in the Responsible Conduct of Research, 2022 (WWW Document) https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-055.html. (Accessed 13 March 2023).
[89] AI Application ChatGPT Temporarily Banned in Italy over Data Collection Concerns, 2023. https://www.cbc.ca/news/world/italy-openai-chatgpt-ban-1.6797963.