

CRC FOCUS



# PRACTICAL GUIDE TO CHIP-SEQ DATA ANALYSIS

Borbala Mifsud  
Kathi Zarnack  
Anaïs F Bardet



A CHAPMAN & HALL BOOK

# Practical Guide to ChIP-seq Data Analysis

## **Focus Computational Biology Series**

This series aims to capture new developments in computational biology and bioinformatics in concise form. It seeks to encourage the rapid and wide dissemination of material for emerging topics and areas that are evolving quickly. The titles included in the series are meant to appeal to students, researchers, and professionals involved in the field. The inclusion of concrete examples and applications, and programming techniques and examples, is highly encouraged.

### **Systems-Level Understanding of Microbial Communities**

Theory and Practice

*Aarthi Ravikrishnan and Karthik Raman*

### **Practical Guide to ChIP-seq Data Analysis**

*Borbala Mifsud, Kathi Zarnack, Anaïs F Bardet*

# Practical Guide to ChIP-seq Data Analysis

Borbala Mifsud

Kathi Zarnack

Anaïs F Bardet



CRC Press

Taylor & Francis Group

Boca Raton London New York

---

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

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

© 2019 by Taylor & Francis Group, LLC  
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed on acid-free paper  
Version Date: 20181008

International Standard Book Number-13: 978-1-138-59652-8 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

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

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

---

# Preface

---

Over the past decade, the experimental ChIP-seq protocol as well as the associated computational analysis methods evolved rapidly. Researchers face a plethora of possible experimental and analytical ChIP-seq approaches, and the lack of a single perfect recipe can make use of the protocol daunting from the start. Each ChIP-seq experiment needs to be tailored to the protein or modification of interest, to the studied organism and to the biological question. Our aim is to summarise the points that need to be considered when performing such a study in order to obtain high-quality and interpretable data. In this book, we will discuss the importance of experimental and analytical choices and give advice for different scenarios. In particular, we will guide the reader through the computational analysis steps from initial quality control through peak calling to downstream analyses and visualisation. This book will thereby show a full workflow, with alternative paths that are suitable for researchers with diverse bioinformatics experience, using Unix command line and R-based solutions.

We hope that this book will help experimental biologists to design their ChIP-seq experiments with the analysis in mind, and to perform the first analysis steps themselves; moreover, it will support bioinformaticians to understand how the data is generated, what the sources of biases are and which methods are appropriate for the different analysis steps.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

---

# Authors

---

**Borbala Mifsud** obtained a PhD in molecular biology at the Institute of Molecular Pathology (IMP) in Vienna, Austria. In the laboratory of Thomas Jenuwein, she worked on epigenetic profiling of a histone methyltransferase mutant mouse, using ChIP-seq. In 2010, she started her postdoctoral work at the EMBL European Bioinformatics Institute (EBI) in the laboratory of Nicholas Luscombe. She is currently an assistant professor at Hamad Bin Khalifa University, Doha, Qatar and honorary lecturer at Queen Mary University London, UK, working on 3D chromatin conformation and the integration of epigenomic data.

**Kathi Zarnack** earned a PhD in molecular biology at the Max-Planck Institute for Terrestrial Microbiology in Marburg, Germany, working on the impact of posttranslational modifications on transcription factor specificity. Moving into bioinformatics, she then joined the EMBL European Bioinformatics Institute (EBI) in Hinxton, UK, as a postdoctoral researcher in the group of Nicholas Luscombe. Since 2014, she leads a research group on Computational RNA Biology at the Buchmann Institute for Molecular Life Sciences (BMLS), Goethe University Frankfurt, Germany.

**Anaïs F. Bardet** completed a PhD in computational biology in the laboratory of Alexander Stark at the Institute of Molecular Pathology (IMP) in Vienna, Austria. She studied the conservation of transcription factor binding sites in different *Drosophila* species and developed tools and pipelines for the comparative analysis of ChIP-seq data. She then worked as a postdoctoral researcher in the laboratory of Dirk Schübeler at the Friedrich Miescher Institute for Biomedical Research (FMI) in Basel, Switzerland, where she investigated the sensitivity of transcription factors to DNA methylation. Since 2017, she is a tenured researcher at the National Center for Scientific Research (CNRS) at the University of Strasbourg, France, where she develops projects exploring the regulation of transcription factor binding.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

---

# Contents

---

CHAPTER	1 ■ Introduction to ChIP-seq	1
BORBALA MIFSUD		
1.1	CHIP-SEQ EXPERIMENT	1
1.2	IMPROVED DETECTION PROTOCOLS	4
1.2.1	ChIP-exo	4
1.2.2	ChIP-nexus	4
1.2.3	CUT&RUN	4
1.2.4	DamID	5
1.3	CHIP-SEQ DATA ANALYSIS WORKFLOW	5
1.4	DESIGNING A CHIP-SEQ EXPERIMENT	6
1.4.1	ChIP-seq controls	6
1.4.2	Sources of bias	6
1.4.3	Antibody quality	7
1.4.4	Read depth	8
1.4.5	Read properties	8
1.4.6	Replicates	9
CHAPTER	2 ■ Getting Started	11
ANAI'S BARDET		
2.1	CHIP-SEQ DATASETS	11
2.2	COMPUTATIONAL REQUIREMENTS	13
2.2.1	Computing environment	13
2.2.2	Data	13
2.2.3	Software	14
2.2.4	File formats	15
2.3	DATA RETRIEVAL FROM GEO	15

x ■ Contents

2.4	CODING TIPS	16
2.5	GRAPHICAL USER INTERFACE TOOLS	19
<b>CHAPTER</b>	<b>3 ■ General Quality Control</b>	<b>21</b>
KATHI ZARNACK		
3.1	INTRODUCTION	21
3.1.1	FASTQ files	21
3.1.2	Available tools	22
3.2	MEASURES OF HTS DATA QUALITY	22
3.2.1	Selected quality metrics	22
3.2.2	FastQC	25
3.3	TRIMMING AND FILTERING	25
3.3.1	Adapter removal	25
3.3.2	Low-quality trimming	26
3.3.3	Trim Galore!	26
<b>CHAPTER</b>	<b>4 ■ Genomic Alignment</b>	<b>27</b>
KATHI ZARNACK		
4.1	INTRODUCTION	27
4.1.1	Alignment concepts	27
4.1.2	Available tools	28
4.2	PARAMETERS AND CONSIDERATIONS	28
4.2.1	Mismatches	28
4.2.2	Multi-mapping	29
4.2.3	Other parameters	29
4.2.4	Output format	31
4.3	GENOMIC ALIGNMENT WITH BOWTIE 2	31
<b>CHAPTER</b>	<b>5 ■ ChIP-seq-specific Quality Control</b>	<b>35</b>
BORBALA MIFSUD		
5.1	CHIP-SEQ-SPECIFIC QUALITY METRICS	35
5.1.1	Signal enrichment	35
5.1.2	Forward and reverse read distribution	38
5.1.3	Duplicate reads	38

5.2 CHIPQC	39
<b>CHAPTER 6 ■ Peak Calling</b>	<b>41</b>
<hr/>	
ANAÏS BARDET	
6.1 CHIP-SEQ SIGNAL TYPES	41
6.1.1 Sharp signal for transcription factors	41
6.1.2 Broad signal for histone marks	42
6.1.3 Mixed signal for RNA polymerase II	42
6.2 GENERAL PEAK CALLING STRATEGY	42
6.2.1 Estimation of fragment size	42
6.2.2 Enrichment of reads	44
6.2.3 Significance score	44
6.2.4 Multiple testing correction	44
6.2.5 Choice of thresholds	44
6.3 EXISTING TOOLS AND CONSIDERATIONS	45
6.3.1 Single-end versus paired-end libraries	45
6.3.2 Sequencing depth and library complexity	46
6.3.3 Experimental resolution	46
6.3.4 New generation of peak callers	47
6.3.5 Post-processing	47
6.4 PEAKZILLA FOR TRANSCRIPTION FACTOR DATA	47
6.5 MACS2 FOR HISTONE MARK DATA	49
6.6 SATURATION ANALYSIS	50
<b>CHAPTER 7 ■ Data Visualisation</b>	<b>53</b>
<hr/>	
ANAÏS BARDET	
7.1 READ DENSITIES	53
7.2 PEAK REGIONS	57
7.3 GENOME BROWSER	57
<b>CHAPTER 8 ■ Comparative Analysis</b>	<b>59</b>
<hr/>	
ANAÏS BARDET AND BORBALA MIFSUD	
8.1 OVERLAP OF PEAK REGIONS	59