Translational Bioinformatics 18
Series Editor: Xiangdong Wang, MD, PhD, Prof

Kang Ning *Editor*

Traditional Chinese Medicine and Diseases

An Omics Big-data Mining Perspective

Springer

Translational Bioinformatics

Volume 18

Series Editor

Xiangdong Wang, Shanghai Institute of Clinical Bioinformatics, Zhongshan Hospital Institute of Clinical Science, Fudan University Shanghai Medical College, Shanghai, China

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Kang Ning
Editor

Traditional Chinese Medicine and Diseases

An Omics Big-data Mining Perspective

Springer

Editor
Kang Ning
College of Life Science and Technology
Huazhong University of Science and Technology
Wuhan, China

ISSN 2213-2775 ISSN 2213-2783 (electronic)
Translational Bioinformatics
ISBN 978-981-19-4770-4 ISBN 978-981-19-4771-1 (eBook)
https://doi.org/10.1007/978-981-19-4771-1

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Traditional Chinese Medicine (TCM) has been in existence for several thousand years. Only in recent tens of years, with the development of sequencing techniques, we have realized that TCM research could be digitalized and standardized with the advancement of the omics approach. This has also raised several anxieties and concerns since so much heterogeneous data has been and could be obtained, and so many questions remain to be answered based on omics data for the TCM research area.

In the omics age, we have faced a paramount of multi-omics data from various sources, which is good since these omics data would provide clues for several biomedical or clinical applications. However, it has also created hurdles for omics data integration, data mining, and in-depth understanding, especially in TCM research.

Several issues are of special interest for studying the omics toward TCM and diseases; these include multi-omics data integration techniques, multi-omics data mining techniques, TCM preparation source tracking, TCM preparation quality control, TCM preparation network pharmacology analysis, as well as various tools and databases supporting the omics researches for TCM and diseases. These topics could be divided into three categories: the basics of omics and data mining techniques for TCM researches, the analysis of biological and chemical components of TCM preparations, as well as network pharmacology analysis of TCM. These constitute a holistic procedure for TCM research, from omics study design to TCM-related omics data integration and data mining, to analyze and interpret the effects of TCM on diseases. However, currently, there is a lack of systematic introduction of such a holistic procedure toward an omics big data mining perspective for TCM and diseases.

In this book, we have featured not only multi-omics big data integration and data mining techniques, but also have focused on cutting-edge omics research techniques and applications for a deep understanding of TCM and diseases. These contents are logically organized in the following order: (1) Basics about multi-omics data and analytical methods for TCM and diseases, (2) the needs of omics studies in TCM researches, and the basic background of omics research in TCM and disease, (3) multi-omics big data integration techniques for TCM researches, (4) multi-omics big data integration and data mining techniques: data integration and data mining techniques will be introduced, with illustrative examples and figures to a better understanding of the essence of the definition of both multi-omics and data mining, as well as how they could be combined for most insights from these omics data, for TCM and disease researches, (5) TCM preparation quality control: using both biological and chemical ingredients, check if any prescribed ingredient is missing, and if any contaminant is present, (6) TCM preparation source tracking: given a sample of unclear origin, determine from which manufacturer and batch the sample is most likely to be from, (7) TCM preparation network pharmacology analysis, (8) TCM analysis data resources, web services, and visualizations, (9) TCM geoherbalism examination and authentic TCM identification.

Devised as a book on TCM and disease researches in the omics age, this book has put a focus on data integration and data mining methods for multi-omics researches, which will be explained in detail and with supportive examples for the "What," "Why," and "How" of omics on TCM-related researches. It attempts to bridge the gap between TCM-related multi-omics big data and data mining techniques for best practices of contemporary bioinformatics and in-depth insights for the TCM-related questions.

Wuhan, China Kang Ning

About the Book

This book focuses on multi-omics big data integration, data mining techniques, and cutting-edge omics researches in principles and applications for a deep understanding of TCM and diseases. It covers a broad spectrum of aspects of TCM as follows: (1) Basics about multi-omics data and analytical methods for TCM and diseases, (2) the needs of omics studies in TCM researches, and the basic background of omics research in TCM and disease, (3) better understanding of multi-omics big data integration techniques, (4) better understanding of multi-omics big data mining techniques, as well as different applications, for most insights from these omics data for TCM and disease researches, (5) TCM preparation quality control for checking both prescribed and unexpected ingredients including biological and chemical ingredients, (6) TCM preparation source tracking, (7) TCM preparation network pharmacology analysis, (8) TCM analysis data resources, web services, and visualizations, (9) TCM geoherbalism examination and authentic TCM identification.

Traditional Chinese Medicine (TCM) has been in existence for several thousands of years. Only in recent tens of years, we have realized that the research on TCM could be profoundly boosted by omics technologies. Devised as a book on TCM and disease researches in the omics age, this book has put the focus on data integration and data mining methods for multi-omics researches, which will be explained in detail and with supportive examples of the "What," "Why," and "How" of omics approach on TCM-related researches. It attempts to bridge the gap between TCM-related multi-omics big data and data mining techniques for best practices of contemporary bioinformatics and in-depth insights for the TCM-related questions.

Contents

Chapter 1
Introduction to Multi-Omics Data and Analytical Methods for TCM and Diseases

Kang Ning

The advances of omics technologies have profound effects on TCM researches, pushing the TCM research towards a digitized, quantitative, and data-driven science.

1.1 The Importance of TCM Researches

The extensive usage of TCM preparations has boosted the current TCM industry to expand continuously (Qu et al. 2019). However, due to the complicated composition, unstandardized quality, and unclear mechanism of efficacy, the internationalization of TCM preparations is facing a grand bottleneck. TCM preparations need quality evaluation and effect assessment based on better pharmacology understanding of TCM composition, in which information integration and "mining" became very important. The systematic analysis of TCM preparations composition, as well as accurate inferring the pharmacological mechanism of TCM in the treatment of diseases, could consequently promote the digitalization and internationalization of TCM preparations (Cheng et al. 2019). However, there is still a lack of data integration and mining methods to decode the complicated but important relationships between TCM and diseases.

K. Ning (✉)
School of Life Science and Technology, Huazhong University of Science and Technology,
Wuhan, Hubei, China
e-mail: ningkang@hust.edu.cn

K. Ning (ed.), *Traditional Chinese Medicine and Diseases*, Translational
Bioinformatics 18, https://doi.org/10.1007/978-981-19-4771-1_1

1.2 Multi-Omics Studies for TCM Researches

The studies for TCM researches rely more and more heavily on multi-omics data integration and analysis, including but not limited to: TCM geoherbalism examination, TCM preparation ingredient identification, TCM source tracking, TCM network pharmacology, *etc*. Among all of these TCM researches, data integration and analysis have acted as the foundation for quantitative evaluation of the TCM quality, as well as for data-driven hypothesis verification (Subramanian et al. 2020).

TCM Medicinal Plant Cultivation and Geoherbalism Examination Either based on wild type or cultivated type medicinal plants, multi-omics are becoming key components for examinations. The multi-omics for research on medicinal plants include: genomics, transcriptomics, proteomics, metabolomics, as well as medicinal plant's soil or leaf metagenomics, all of which would be essential for investigating the regulation mechanism of the environmental factors in shaping the medicinal plant and generating the effective small molecules of the TCM medicinal plant (Xin et al. 2019).

While China has abundant TCM resources, there is still a shortage of wild Chinese herbal medicine, and there are still certain quality differences between cultivated and wild TCM medicinal plants (Shen et al. 2016). It is about ensuring the controllability and yield of cultivation, as well as restoring the planting environment and ensuring the efficacy of Chinese medicinal plants, that we need to better understand the cultivation of medicinal plants. And such better understandings could be gained through multi-omics investigations.

The integration of multi-omics for medicinal plants, including genomics, transcriptomics, microbiomics, and metabolomics, could strongly deeper our understanding of these medicinal plants (Liu et al. 2017). In medicinal plant genomics and transcriptomics researches, the use of high-throughput sequencing technologies and species genetic information enables us to conduct large-scale excavation and identification of functional genes, thereby advancing the field of TCM geoherbalism research at the molecular level. In medicinal plant microbiomics and metabolomics researches, the effects of environmental factors on medicinal plant's soil or leaf microbial communities (Koberl et al. 2013), as well as on medicinal plant's metabolites (Chamkhi et al. 2021; Mishra et al. 2021). Thus, by combining the multi-omics of medicinal plants, we could not only gain information about how medicinal plants grow under environmental stress, but also how various internal factors in concert generate the metabolites that we have observed for the medicinal plants. Collectively, better understanding of multi-omics could be gained for medicinal plants, and could help for better cultivation of medicinal plants.

TCM Preparation Biological and Chemical Ingredient Identification Both biological and chemical ingredients of TCM preparations are important for the quality of TCM preparation.

The original "pills, powders, ointments, and pellets" forms of TCM preparations has evolved into a range of current dosage forms. And its quality control has steadily developed into the four key ways of TCM identification: basic origin, properties, microscopic, and physicochemical (Feng et al. 2013). Quality evaluation methods of TCM should include chemical composition analysis (chemical model test) and biological composition identification (molecular identification) (Bai et al. 2020).

Chemical component analysis is the primary mode of evaluating the quality of TCM preparations (Yao et al. 2021), largely due to its objectivity and digital advantages. Chemical quality control procedures are comprised of commonly used methods for chromatography and spectral analysis, these methods are simple to operate and widely used in natural medicine quality analysis. However, chemical analysis ignores the complex compositional characteristics of Chinese medicinal materials. A TCM preparation usually contains several Chinese medicinal materials, each corresponding to a medicinal plant or chemical mineral. Moreover, each medicinal plant also contains many active ingredients and impurity compositions, many of which may be similar qualitatively or quantitatively. Therefore, analyzing the chemical composition alone is not enough if we aim to fully understand the compositions of TCM preparations.

Biological component analysis quality control is another mode of evaluating the quality of TCM preparations (Cheng et al. 2014). TCM preparation is a mixed system containing multiple biological species (mixed biological samples), the biological composition analysis is essentially the identification of species. Metagenomic research based on high-throughput sequencing technology is one of the most successful and reliable approaches for comprehending and analyzing the composition and function of mixed systems. The application of metagenomics could help in the establishment of a method for evaluating TCM preparation.

TCM Source Tracking TCM safety and quality issues remain a hotspot and focus of attention at home and abroad, if effective assurances cannot be obtained, TCM preparations will be hindered in its international promotion use, resulting in a reduction in TCM's international influence and competitiveness. The traceability of Chinese medicinal materials is usually examined by DNA barcoding technology, resulting in accurate source tracking of medicinal plants and animals against counterfeit materials (Noh et al. 2021). However, few works have been conducted on the traceability of Chinese medicinal preparations (Wang et al. 2020). Recently, machine learning methods have also been proposed for TCM source tracking. These emerging methods have enabled accurate and fast source tracking for a broad-spectrum of TCM preparations.

TCM Network Pharmacology and Disease Association Studies In the omics age, the effects of TCM preparations on treating diseases are analyzed mostly by network pharmacology approach (Tao et al. 2016).

For TCM preparations, it is vital to analyze pharmacodynamic material, metabolic pathway and regulation mechanism *in vivo (*Zhou et al. 2020a*)*. The multi-omics data of a TCM preparations contain multi-components, has multi-links, and has multi-targets, which is asking for a systematic examination of TCM preparations, which is promising for significantly increasing the research efficiency and effectiveness of TCM analyses.

Modern "omics" theory has advanced rapidly in recent years, and the systems biology and bioinformatics tools are being employed to network pharmacology research. The intervention and influence of medicine on a broad spectrum of diseases are systematically studied using network approach (Boezio et al. 2017), revealing the enigma of multi-molecular drugs' synergistic effect on the human body (Wu et al. 2018). Network pharmacology brings new opportunities for the study of TCM preparations, which has the advantage to help us to explain TCM and biological system from network and system perspective (Boezio et al. 2017). Understanding the molecular mechanism of diseases and the pharmaceutical action mechanism from the structure and function of the "biomolecular network" is one of the key processes. And using this process, the research mode for understanding TCM preparations effectiveness is shifting from descriptive to predictive research (Miryala et al. 2018).

Another approach for systematic assessment of TCM preparations and their associations with diseases is the association study (Wu et al. 2019). The association study has focused on two key objectives: biomarker selection, and linear or non-linear association model building. In TCM preparation analysis (Bai et al. 2019), the association analysis could not only provides valuable information about which components have contributed to the treatment of diseases quantitatively, but also could help for better prediction of the effects of TCM preparations on certain diseases (Chen et al. 2019).

1.3 Representative Studies

Several studies have been conducted for TCM and disease researches, focusing on multi-omics data integration and data mining. These studies include but are not limited to: Biological ingredient analysis for several representative TCM preparations, TCM network pharmacology for treating inflammation diseases, *etc*. Here are four representative studies that use multi-omics data and analytical methods for TCM and disease researches.

Analysis of Biological Components of Da Huoluo Wan (DHW) (Yao et al. 2022) DHW samples were bought from two different manufacturers (marked as A and B) with three batches (I, II and III) Then, these samples were conducted with three biological replicates and then were amplified with ITS2 and *trnL*, respectively. Through high-sequencing technology, all ingredients in the DHW were detected, including 38 prescribed herbal species (PHS), covering 28 prescribed herbal materials (PHMs), and other species (Fig. 1.1). Based on ITS2, 35 PHS covering 25 PHMs (Fig. 1.1a), including the processed herbal materials such as the stir-fried Baishu were detected, while 18 PHMs and 22 PHS were detected based on *trnL* (Fig. 1.1b). The species compositions of products from different batches and manufacturers of the same TCM are almost the same, but the relative abundances are different.

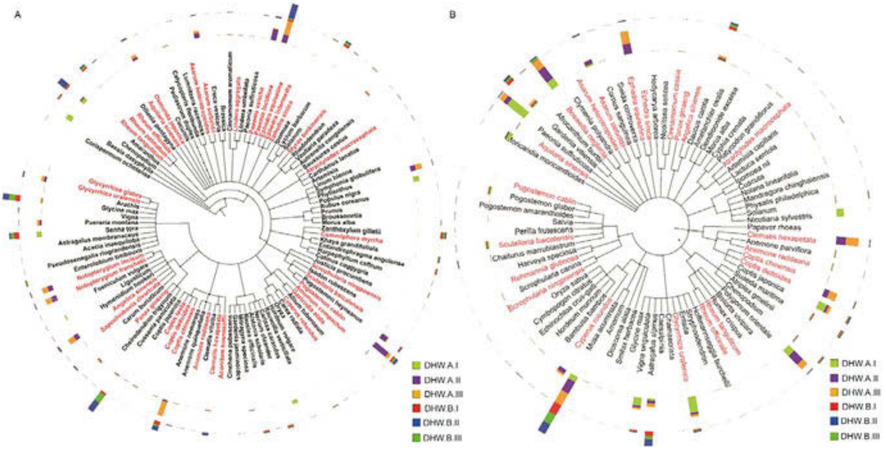

Fig. 1.1 Phylogenetic analysis of the representative species that had at least 0.1% relative abundance in DHW samples (Yao et al. 2022). (**a**) Based on ITS2; (**b**) Based on *trnL*. The branch in the tree depicts the taxonomic classification of species. The word marked in red means the prescribed herbal species, and the colorful bar means the average relative abundance of species across the three batches from the two manufacturers (**a**, **b**). **(Reprinted with permission from authors of** Yao et al. (2022))

Analysis of the Effects of Water Extract of Ganodermalucidum Mycelium on Treating HFD-Fed Mice Mediated by Gut Microbiome (Chang et al. 2017) Data from Chih-Jung Chang et al. show that *Ganoderma lucidum* water extract (WEGL) not only reverses HFD-induced gut dysbiosis (as evidenced by lower Firmicutes-to-Bacteroidetes ratios and endotoxin-bearing Proteobacteria levels), but also maintains intestinal barrier integrity and reduces metabolic endotoxemia. The anti-obesity and microbiota-modulating effects are transmitted via horizontal feces transfer from WEGL-treated animals to HFD-fed mice. First, WEGL reduced body weight, inflammatory response, and insulin resistance in HFD-induced obese mice, and WEGL can reduce the ratio of Firmicutes to Bacteroidetes in the intestinal flora and the level of endotoxin-containing *Proteus*, maintain the integrity of the intestinal barrier, and reduce metabolic endotoxemia. Furthermore, transplanting WEGL-interfered feces into HFD mice lowered obesity, decreased the amount of pro-inflammatory substances, raised the expression of tight junction proteins Occludin and ZO-1, and alleviated intestinal flora imbalance. WEGL's high molecular weight polysaccharide of more than 300 kDa is the key to its potency.

Analysis of the Effects of *Cordyceps Sinensis* on Treating HFD-Fed Mice Mediated by Gut Microbiome (Chen et al. 2017) *Cordyceps sinensis* mycelia and its high molecular weight polysaccharide components have been shown to promote the growth of specific intestinal bacteria, such as Parabacteroides guesei, which improves obesity and related metabolic disorders in mice, according to