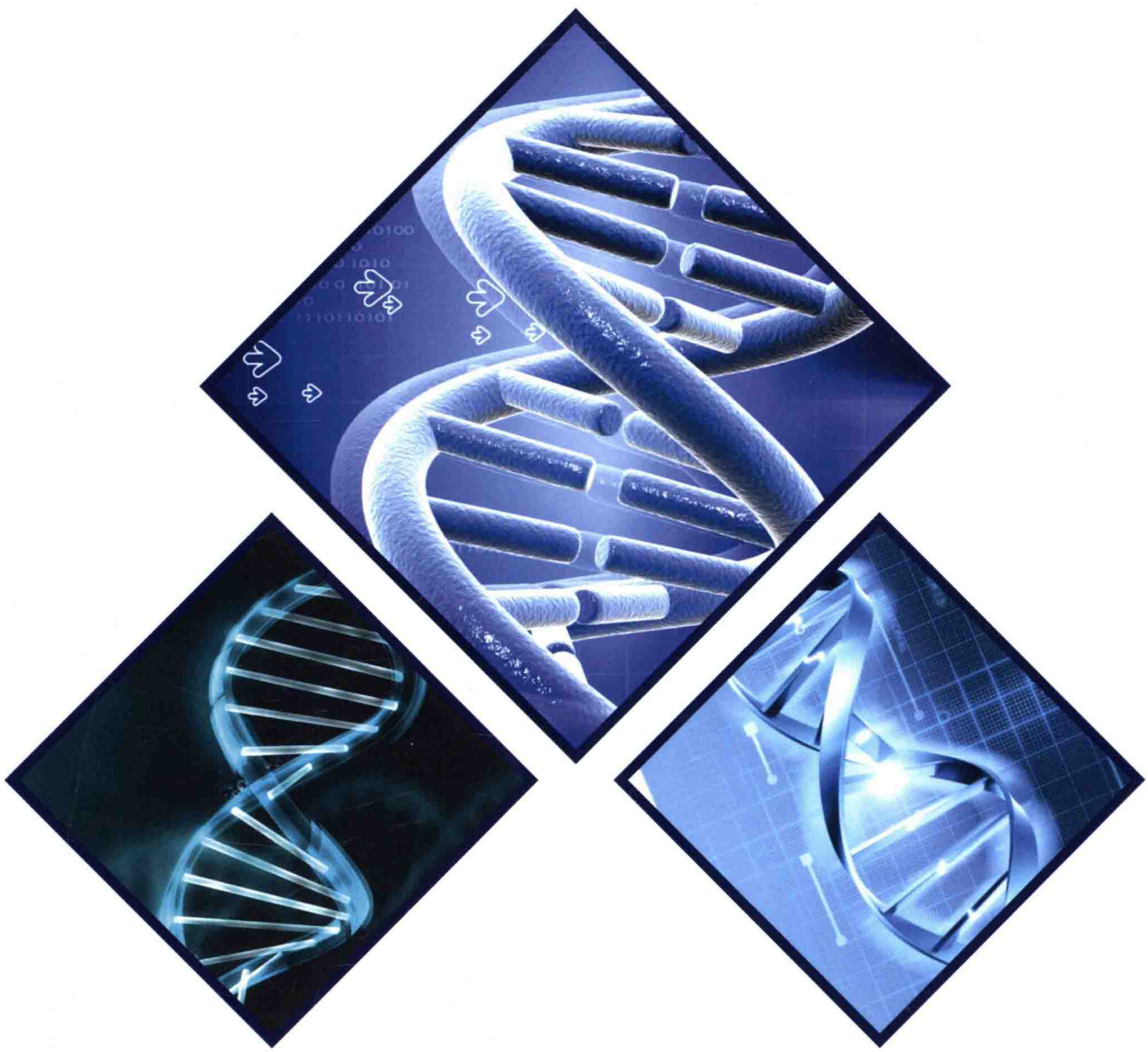
Principles of **Computational Biology and Genome Analysis**



Daniel McGuire

Principles of Computational Biology and Genome Analysis

About the Book

Genome analysis has changed the way biological and anthropological evolution has been perceived. Computational analysis of genetic data has made it possible for the creation of speculative models that can predict possible evolutionary patterns while taking into account natural biological phenomena such as aging, disease and degeneration of the body. This book on computational biology and genome analysis contributes to the fields of computational neuroscience and computational evolutionary biology. The various studies that are constantly contributing towards advancing technologies and evolution of this field are examined in detail in this text. It elucidates new techniques and their applications in a multidisciplinary approach. This book is a vital tool for all researching or studying computational biology and genome analysis as it gives incredible insights into emerging trends and concepts.

About the Editor

Daniel McGuire holds a PhD in computational biology & bioinformatics from the University of Southern California, United States. He is an associate professor of computational biology in a leading university of America. His academic and research interests lie in bioinformatics algorithms and computational cell biology. McGuire has organized many national conferences on computational biology and bioinformatics; and has authored and edited a number of articles, research papers and book chapters in these fields.

ISBN 978-1-68286-407-4

SYRAWOOD
PUBLISHING HOUSE
www.syrawoodpublishinghouse.com

McGuire

Principles of Computational Biology and Genome Analysis

SYRAWOOD PUBLISHING HOUSE

Principles of Computational Biology and Genome Analysis

Edited by Daniel McGuire

SYRAWOOD
PUBLISHING HOUSE
New York

Published by Syrawood Publishing House,
750 Third Avenue, 9th Floor,
New York, NY 10017, USA
www.syrawoodpublishinghouse.com

Principles of Computational Biology and Genome Analysis
Edited by Daniel McGuire

© 2017 Syrawood Publishing House

International Standard Book Number: 978-1-68286-407-4 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Copyright for all individual chapters remain with the respective authors as indicated. All chapters are published with permission under the Creative Commons Attribution License or equivalent. A wide variety of references are listed. Permission and sources are indicated; for detailed attributions, please refer to the permissions page and list of contributors. Reasonable efforts have been made to publish reliable data and information, but the authors, editors and publisher cannot assume any responsibility for the validity of all materials or the consequences of their use.

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy. Furthermore, the publisher ensures that the text paper and cover boards used have met acceptable environmental accreditation standards.

Trademark Notice: Registered trademark of products or corporate names are used only for explanation and identification without intent to infringe.

Cataloging-in-publication Data

Principles of computational biology and genome analysis / edited by Daniel McGuire.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-68286-407-4
1. Computational biology. 2. Genomes. 3. Bioinformatics. I. McGuire, Daniel.
QH324.2. P74 2017
570.28--dc23

Printed in China.

此为试读,需要完整PDF请访问: www.ertongbook.com

Principles of Computational Biology and Genome Analysis

PREFACE

Genome analysis has changed the way biological and anthropological evolution has been perceived. Computational analysis of genetic data has made it possible for the creation of speculative models that can predict possible evolutionary patterns while taking into account natural biological phenomena such as aging, disease and degeneration of the body. This book on computational biology and genome analysis contributes to the fields of computational neuroscience and computational evolutionary biology. The various studies that are constantly contributing towards advancing technologies and evolution of this field are examined in detail in this text. It elucidates new techniques and their applications in a multidisciplinary approach. This book is a vital tool for all researching or studying computational biology and genome analysis as it gives incredible insights into emerging trends and concepts.

The purpose of the book is to provide a glimpse into the dynamics and to present opinions and studies of some of the scientists engaged in the development of new ideas in the field from very different standpoints. This book will prove useful to students and researchers owing to its high content quality.

At the end, I would like to appreciate all the efforts made by the authors in completing their chapters professionally. I express my deepest gratitude to all of them for contributing to this book by sharing their valuable works. A special thanks to my family and friends for their constant support in this journey.

Editor

TABLE OF CONTENTS

Preface IX

Chapter 1 **Somatic Mutations Favorable to Patient Survival are Predominant in Ovarian Carcinomas** 1
Wensheng Zhang, Andrea Edwards, Erik Flemington, Kun Zhang

Chapter 2 **Development of a Gene-Centered SSR Atlas as a Resource for Papaya (*Carica papaya*) Marker-Assisted Selection and Population Genetic Studies** 13
Newton Medeiros Vidal, Ana Laura Grazziotin, Helaine Christine Cancela Ramos, Messias Gonzaga Pereira, Thiago Motta Venancio

Chapter 3 **The Genome of the Generalist Plant Pathogen *Fusarium avenaceum* is Enriched with Genes Involved in Redox, Signaling and Secondary Metabolism** 23
Erik Lysøe, Linda J. Harris, Sean Walkowiak, Rajagopal Subramaniam, Hege H. Divon, Even S. Riiser, Carlos Llorens, Toni Gabaldón, H. Corby Kistler, Wilfried Jonkers, Anna-Karin Kolseth, Kristian F. Nielsen, Ulf Thrane, Rasmus J. N. Frandsen

Chapter 4 **Sequencing, Annotation and Analysis of the Syrian Hamster (*Mesocricetus auratus*) Transcriptome** 40
Nicolas Tchitchek, David Safronetz, Angela L. Rasmussen, Craig Martens, Kimmo Virtaneva, Stephen F. Porcella, Heinz Feldmann, Hideki Ebihara, Michael G. Katze

Chapter 5 **Marked Microevolution of a Unique *Mycobacterium tuberculosis* Strain in 17 Years of Ongoing Transmission in a High Risk Population** 51
Carolina Mehaffy, Jennifer L. Guthrie, David C. Alexander, Rebecca Stuart, Elizabeth Rea, Frances B. Jamieson

Chapter 6 **Comparative Analysis of Predicted Plastid-Targeted Proteomes of Sequenced Higher Plant Genomes** 61
Scott Schaeffer, Artemus Harper, Rajani Raja, Pankaj Jaiswal, Amit Dhingra

Chapter 7 ***De novo* Assembly of the Grass Carp *Ctenopharyngodon idella* Transcriptome to Identify miRNA Targets Associated with Motile Aeromonad Septicemia** 81
Xiaoyan Xu, Yubang Shen, Jianjun Fu, Liqun Lu, Jiale Li

Chapter 8 **Uromodulin Retention in Thick Ascending Limb of Henle's Loop Affects SCD1 in Neighboring Proximal Tubule: Renal Transcriptome Studies in Mouse Models of Uromodulin-Associated Kidney Disease** 91
Marion Horsch, Johannes Beckers, Helmut Fuchs, Valérie Gailus-Durner, Martin Hrabě de Angelis, Birgit Rathkolb, Eckhard Wolf, Bernhard Aigner, Elisabeth Kemter

Chapter 9 **A Bumpy Ride on the Diagnostic Bench of Massive Parallel Sequencing, the Case of the Mitochondrial Genome** 101
Kim Vancampenhout, Ben Caljon, Claudia Spits, Katrien Stouffs, An Jonckheere, Linda De Meirleir, Willy Lissens, Arnaud Vanlander, Joél Smet, Boel De Paepe, Rudy Van Coster, Sara Seneca

Chapter 10 **Zinc-Finger Nuclease Knockout of Dual-Specificity Protein Phosphatase-5 Enhances the Myogenic Response and Autoregulation of Cerebral Blood Flow in FHH.1^{BN} Rats** 110
Fan Fan, Aron M. Geurts, Mallikarjuna R. Pabbidi, Stanley V. Smith, David R. Harder, Howard Jacob, Richard J. Roman

Chapter 11 **Disclosure of Genetic Information and Change in Dietary Intake: A Randomized Controlled Trial** 120
Daiva E. Nielsen, Ahmed El-Sohemy

Chapter 12 **A Simplified and Versatile System for the Simultaneous Expression of Multiple siRNAs in Mammalian Cells using Gibson DNA Assembly** 128
Fang Deng, Xiang Chen, Zhan Liao, Zhengjian Yan, Zhongliang Wang, Youlin Deng, Qian Zhang, Zhonglin Zhang, Jixing Ye, Min Qiao, Ruifang Li, Sahitya Denduluri, Jing Wang, Qiang Wei, Melissa Li, Nisha Geng, Lianggong Zhao, Guolin Zhou, Penghui Zhang, Hue H. Luu, Rex C. Haydon, Russell R. Reid, Tian Yang, Tong-Chuan He

Chapter 13 **Comparison of the Transcriptome between Two Cotton Lines of Different Fiber Color and Quality** 140
Wenfang Gong, Shoupu He, Jiahuan Tian, Junling Sun, Zhaoe Pan, Yinhua Jia, Gaofei Sun, Xiongming Du

Chapter 14 **Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement** 152
Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, Ashlee M. Earl

Chapter 15 **Complete Chloroplast Genome Sequence of Omani Lime (*Citrus aurantiifolia*) and Comparative Analysis within the Rosids** 166
Huei-Jiun Su, Saskia A. Hogenhout, Abdullah M. Al-Sadi, Chih-Horng Kuo

Chapter 16 **IL28B Polymorphism cannot Predict Response to Interferon Alpha Treatment in Patients with Melanoma** 177
Martin Probst, Christoph Hoeller, Peter Ferenci, Albert F. Staettermayer, Sandra Beinhardt, Hubert Pehamberger, Harald Kittler, Katharina Grabmeier-Pfistershammer

Chapter 17 **Identification of Allelic Heterogeneity at Type-2 Diabetes Loci and Impact on Prediction** 184
Yann C. Klimentidis, Jin Zhou, Nathan E. Wineinger

Chapter 18 **Comparative Transcriptomic Analysis of the Response to Cold Acclimation in *Eucalyptus dunnii*** 191
Yiqing Liu, Yusong Jiang, Jianbin Lan, Yong Zou, Junping Gao

此为试读,需要完整PDF请访问: www.ertongbook.com

Chapter 19 **Vitamin D Insufficiency in Arabs and South Asians Positively Associates with Polymorphisms in GC and CYP2R1 Genes** 202
Naser Elkum, Fadi Alkayal, Fiona Noronha, Maisa M. Ali, Motasem Melhem, Monira Al-Arouj, Abdullah Bennakhi, Kazem Behbehani, Osama Alsmadi, Jehad Abubaker

Chapter 20 **Genome-Wide Investigation and Expression Profiling of AP2/ERF Transcription Factor Superfamily in Foxtail Millet (*Setaria italica L.*)** 210
Charu Lata, Awdhesh Kumar Mishra, Mehanathan Muthamilarasan, Venkata Suresh Bonthala, Yusuf Khan, Manoj Prasad

Chapter 21 **Permanent Cardiac Sarcomere Changes in a Rabbit Model of Intrauterine Growth Restriction** 224
Iratxe Torre, Anna González-Tendero, Patricia García-Cañadilla, Fátima Crispi, Francisco García-García, Bart Bijnens, Igor Iruretagoyena, Joaquin Dopazo, Ivan Amat-Roldán, Eduard Gratacós

Permissions

List of Contributors

Index

Somatic Mutations Favorable to Patient Survival Are Predominant in Ovarian Carcinomas

Wensheng Zhang[1], Andrea Edwards[1], Erik Flemington[2]*, Kun Zhang[1]*

1 Department of Computer Science, Xavier University of Louisiana, New Orleans, Louisiana, United States of America, **2** Tulane Cancer Center, Tulane School of Medicine, New Orleans, Louisiana, United States of America

Abstract

Somatic mutation accumulation is a major cause of abnormal cell growth. However, some mutations in cancer cells may be deleterious to the survival and proliferation of the cancer cells, thus offering a protective effect to the patients. We investigated this hypothesis via a unique analysis of the clinical and somatic mutation datasets of ovarian carcinomas published by the Cancer Genome Atlas. We defined and screened 562 macro mutation signatures (MMSs) for their associations with the overall survival of 320 ovarian cancer patients. Each MMS measures the number of mutations present on the member genes (except for TP53) covered by a specific Gene Ontology (GO) term in each tumor. We found that somatic mutations favorable to the patient survival are predominant in ovarian carcinomas compared to those indicating poor clinical outcomes. Specially, we identified 19 (3) predictive MMSs that are, usually by a nonlinear dose-dependent effect, associated with good (poor) patient survival. The false discovery rate for the 19 "positive" predictors is at the level of 0.15. The GO terms corresponding to these MMSs include "lysosomal membrane" and "response to hypoxia", each of which is relevant to the progression and therapy of cancer. Using these MMSs as features, we established a classification tree model which can effectively partition the training samples into three prognosis groups regarding the survival time. We validated this model on an independent dataset of the same disease (Log-rank p-value $<2.3\times10^{-4}$) and a dataset of breast cancer (Log-rank p-value $<9.3\times10^{-3}$). We compared the GO terms corresponding to these MMSs and those enriched with expression-based predictive genes. The analysis showed that the GO term pairs with large similarity are mainly pertinent to the proteins located on the cell organelles responsible for material transport and waste disposal, suggesting the crucial role of these proteins in cancer mortality.

Editor: Shannon M. Hawkins, Baylor College of Medicine, United States of America

Funding: Research reported in this publication was supported by a National Institutes of Health grant (NIGMS-2G12MD007595), an US Department of Army grant (W911NF-12-1-0066) and a seed grant from the Louisiana Cancer Research Consortium. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: eflemin@tulane.edu (EF); kzhang@xula.edu (KZ)

Introduction

Ovarian cancer is the fifth-leading cause of cancer death among women in the United States [1]. The disease is often called a "silent killer" since its occurrence is usually not detected until an advanced stage. About 70% of the deaths occur in patients with advanced-stage, high-grade serous ovarian carcinomas [2]. The mortality has not been significantly improved in the past three decades [3]. Except for the detection delay and inaccessible location of the ovaries, other factors accounting for the persistent mortality include the poor understanding of the underlying biology and a lack of reliable biomarkers [4].

The formation of tumors largely results from cell growth that gets out of control [5]. In the human genome, there are many different types of genes that control cell growth in a very systematic, precise way. When these genes have an error in their DNA codes, the RNA or proteins that they encode may not function properly. Typically, a series of several mutations to certain classes of genes is usually required before a normal cell will transform into a cancer cell [6]. Nevertheless, some observed mutations may be neutral or even beneficial to patient survival. This perception can be considered from at least two perspectives. First, some mutations may be deleterious to the growth and proliferation of cancer cells, thus offering a protective mechanism to the patients. Second, some mutations may include the actual causal factors for relatively less-malignant subtypes of the same disease. For example, previous studies showed that cases with BRCA1/2 mutations have better overall survival than those with wild type BRCA1/2 in patients with ovarian carcinoma [7,8].

To date, the Cancer Genome Atlas (TCGA) [9] has generated and released comprehensive genomic, epigenomic and proteomic data of clinically annotated high-grade serous ovarian carcinomas (Ov-HGSCs). These rich data provide an unprecedented opportunity to investigate the genetic mechanisms underlying the variance in the survival of cancer patients and to advance the clinical prognosis and therapy of the disease. Besides the BRCA1/2 genotypes, the TCGA ovarian cancer paper [7] showed that gene expression-based sample clusters are also associated with the survival outcomes. Moreover, recent years have witnessed numerous studies that focus on the re-analysis of the TCGA data.

In these works, miscellaneous predictive signatures for survival outcomes have been identified. These signatures include the expression measures of coding and miRNA genes [10], genotypes of germline single-nucleotide polymorphisms (SNPs) [11], methylation patterns of genes in key cancer pathways [12], DNA copy number variations (CNV) [13] and the occurrences of chromosome aberrations [14].

As shown in [7], most of the Ov-HGSCs had 8 to 209 somatic mutations. These mutations, detected by exome sequencing, were present in 8945 genes, and 92% of them were validated by experiments using alternative technologies. However, most of the observed variants may be passenger mutations not involved in the formation and progression of ovarian cancer. Hidden among observed mutations are the individual-specific tumor drivers and the genetic alterations positively or adversely impacting the growth and survival of cancer cells. The identification of the clinically important mutations (genes) is far from completed. A major challenge impeding the effective statistical analysis of the somatic mutation spectrum (SMS) is the data sparseness issue. This is particularly implied by the fact that, among the 510 consensus cancer genes collected in the Catalogue Of Somatic Mutations In Cancer database [15], only six are significant in terms of their mutation frequencies over the 326 tumors. Nevertheless, two recent studies have demonstrated the potential to train a predictive model for survival outcomes of ovarian cancer patients using SMS [16,17]. In this study, we conducted a unique analysis of the recently updated TCGA's clinical and SMS datasets of ovarian cancer. Our study provides significant insights into the treatment of ovarian cancer and may open novel avenues for molecular prognosis and prediction.

Results

Predictive macro mutation signatures for patient survival

We developed a novel method to unravel the relationships between the somatic mutations and the survival time of cancer patients. First, by assuming that the DNA alterations on the genes of a similar function may have equivalent or complementary impacts on the growth and proliferation of cancer cells, we defined 562 macro mutation signatures (MMS), each of which corresponds to a highly-specific Gene Ontology (GO) term with 50 to 500 member genes. For each patient (i.e. a carcinoma sample), the MMS quantities were calculated as the number of the mutations on the genes (except for TP53) covered by the cognate GO term. When a gene involves in multiple GO terms, the mutation(s) present on each gene were counted with respect to each cognate MMS. In this way, we circumvented the sparsity issue inherent to the raw somatic mutation data (see Introduction section). After that, the MMSs were screened for their associations with the overall survival (OS) months of the cancer patients. More specifically, the associations were evaluated by performing the Log-rank test and Cox Proportional Hazards (Cox-PH) regression analysis on the mutation and clinical datasets of 320 training samples. In the implementation, quantities of the MMSs were capped by a ceiling value of 2, which represented that a tumor had at least two mutations present on the member genes covered by the corresponding GO term. Capping the MMS values was performed to alleviate the influence of leverage data points, which were related to un-ordinarily high MMS values and usually occurred in highly-specific GO terms. In the Cox-PH analysis, along with a focused MMS, the ages of the patients at the initial diagnosis and a binary measurement variable indicating the presence of somatic mutation on TP53 gene, which had a modestly significant ($p<0.05$) effect on the patient survival as shown in our preliminary analysis of the same data, were included as covariates. In the Log-rank test, the three possible values (0, 1, 2) of a specific MMS were factorized as the indicators of three groups.

The analysis of the training set ($N = 320$) demonstrated strong evidence for the existence of an association between the MMSs and survival outcomes. As shown in Figure 1-A and 1-B, the distributional profiles of the p-values (from both the Log-rank test and Cox-PH regression) for the MMSs are deviated from a uniform distribution U (0, 1). Interestingly, most of the regression coefficients (i.e., beta values), especially those corresponding to small p-values, are negative (Figure 1-C). In the Cox-PH model, a negative regression coefficient indicates that the hazard function decreases (or equivalently, survival time increases) as the quantity of the corresponding predictive variable increases [18]. In this regard, we concluded that somatic mutations favorable to the survival of cancer patients are predominant in ovarian carcinoma compared to those indicating poor clinical outcomes. As shown in Figure 1-D and 1-E, this statement is also valid in terms of the number of the involved GO terms and the sizes of the relevant gene sets (Table S1).

Neither Log-rank test nor Cox-PH regression analysis are perfect for evaluating the associations between a MMS and the clinical outcome. The former ignores the patients' ages at the initial diagnosis, which intuitively influence survival time. The latter assumes that the quantity of the hazard functions is linearly dependent on the preprocessed MMS values, which is not true in many cases. Therefore, we determined the top significant MMSs (GO terms) by an alternative method. That is, we selected 20 MMSs if (1) their p-values from both the Log-rank test and Cox-PH analysis are less than 0.05 and (2) the resulting composite p-value (see Method section) is less than 0.025. Among those MMSs, 19 are "positive" predictors for survival time. Considering that the selection criteria may be too stringent for the potential MMSs adversely affecting overall survival outcomes, we chose another two MMSs. These two "negative" predictors, with Log-rank p-values less than 0.01, are relevant to two small patient sets of poor survival and correspond to GO:0045666 and GO:0042393, respectively. In this way, we established a predictor set consisting of 22 MMSs (Table 1).

We addressed the multiple-testing problem in the identification of predictive MMSs for patient survival by calculating false discovery rate (FDR) with a permutation-based algorithm (see Material and Methods section). In the implementation, we considered not only the skewness of the distribution of the effect parameters estimated from the original datasets (Figure 1C), but also the asymmetry of their null distribution established from the randomly permutated datasets (Figure S1). Because only one negative predictor (MMS) was rigorously selected, the analysis was focused on the 19 MMSs associated with good clinical outcomes. The result showed that, when those MMSs are stated to be significant, the FDR could be controlled at the level of 0.15.

Based on the definitions, we partition the cognate GO terms of the 19 positive predictors (MMSs) into six groups: (1) the gene products (proteins) locate in the cell organelles (membrane) responsible for waste disposal (lysosome) and material transport (recycling endosome); (2) the gene products locate in the subcellular structures playing roles in mitosis (nucleosome, spindle pole, centrosome); (3) the gene products perform function in cell division (mitosis and cytokinesis); (4) the gene products are involved in cellular responses to nutrient and hypoxia; (5) the gene products play roles in cancer pathways (integrin-mediated signaling pathway and positive regulation of ERK1 and ERK2 cascade); and (6) others. Numerous records regarding those GO

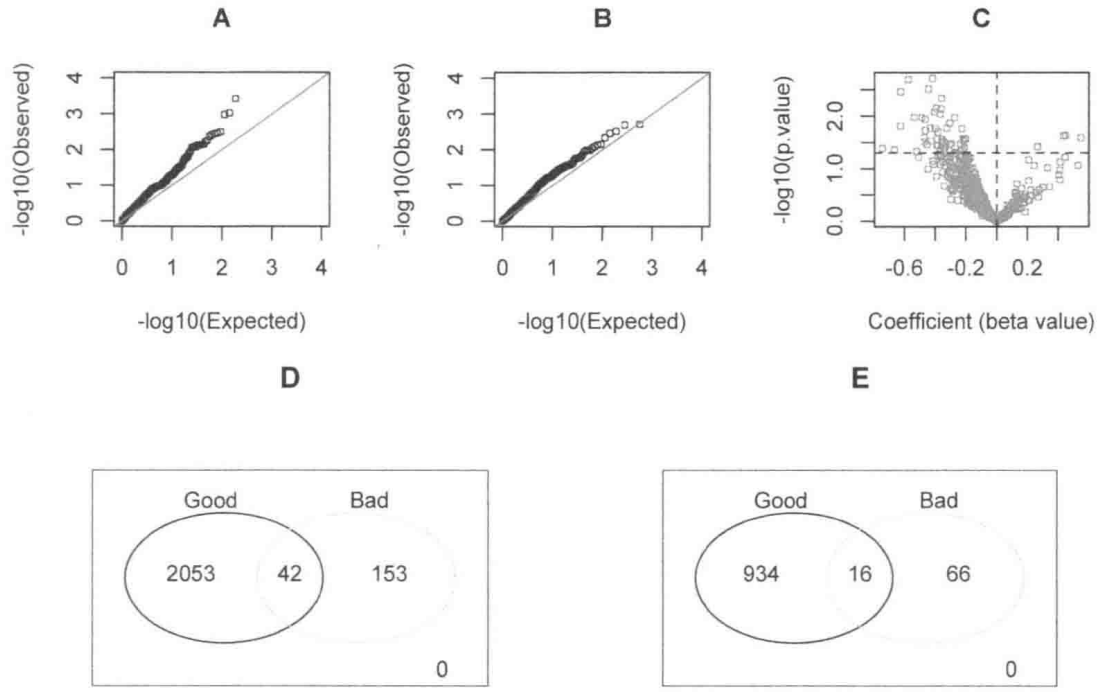


Figure 1. The profile for the associations between the somatic mutations and survival time of patients with ovarian cancer. A (B): The Q-Q plot of the p-values from Log-rank test (Cox-PH regression) for the 562 considered MMSs. C: The volcano plot of the Cox-PH p-values and regression coefficients for the 562 considered MMSs. The horizontal dot line marks p = 0.05. D: The Venn diagram for the entire set of genes covered by the 22 selected MMSs. Specifically, the good (bad) genes are the genes involved in the GO terms corresponding to the 19 (3) positive (negative) MMSs which predict good (poor) clinical outcomes. A gene can belong to both the positive and negative MMSs, therefore may be double counted. E: The Venn diagram for the subset of the genes which are covered by the 22 selected MMSs. Each of the genes has the mutation burden in at least one training sample.

terms' relevance to the formation, progression and therapy of tumors can be located in the literature (see Discussion section).

By looking into the Kaplan-Meier survival curves, we found a "dose-effect" relationship between the somatic mutations and survival outcomes. That is, for a specific GO term, a single mutation on the member genes usually does not make much difference to the patient survival time but two or multiple mutations do (Figure 2).

Robustness analysis of the selected predictors

In order to test the robustness of our main result, we randomly split the 320 training samples into two equal-size subsets, and estimated the effects of the 22 predictive signatures on each subset separately. The result showed that the sign (positive or negative) of the estimated regression coefficients of hazard functions on the MMSs were consistent with those estimated using the entire training set (upper left plot of Figure S2).

We further tested if each of the 22 selected MMSs can individually predict the survival of cancer patients in the validation set. The results showed that, just the predictors ranked at the second, third and fifth places had a marginally significant p-value (upper right and bottom plots of Figure S2). While this analysis only provided a minor support to our findings in the last subsection, the result is aligned to our expectation. This is because the insufficiencies of the training set, i.e. the small sample size (N = 140) and un-validated mutations, could lead to a lower statistical power.

A classification tree model for patient survival prediction

The findings presented above inspired us to build a classification tree to predict the patient survival using the 22 identified MMSs. More specifically, based on the measures of all three negative predictors, we can separate a poor-prognosis group from the entire set of training samples whose members meet $\max_j(N_j^i) \geq 2$, where N_j^i indicates the value of the j^{th} negative MMS on the i^{th} sample. Then, based on the values of the top k positive predictors and the same criterion, i.e. $\max_k(P_k^i) \geq 2$ with P_k^i indicating the value of the j^{th} positive MMS on the i^{th} sample, a good-prognosis group can be split from the remaining samples that constitute an intermediate-prognosis group. See Figure S3 for an illustration. The threshold for the combined MMS values in the partition was heuristically chosen according to the pattern of dose-dependent effect of several MMSs of high interest to patient survival, as showed in Figure 2 and described in the ending paragraph of the first result subsection.

As shown in Figure 3, the patient groups generated by the tree model are significantly differentiated with respect to the times of overall survival (OS) and progression-free survival (PFS). Regardless of the k value (5 or 10), the Log-rank test p-value is less than 1.2×10^{-10} for OS and is less than 1.6×10^{-7} for PFS. From the Kaplan-Meier survival curves, we found that, for the poor-prognosis group, the upper limits of OS and PFS are 50 and 20 months, respectively. They are also the time points when the differences in the survival probabilities between the good-prognosis group and intermediate prognosis group become sharper. It is worth noting that, the choice of k value is somewhat arbitrary. The value determines the size of the predicted good-prognosis group that has a better survival curve compared to the intermediate-prognosis group. Therefore, a prior knowledge about the proportion of good-prognosis samples would help with the specification of k value.

Table 1. The summary of significant MMSs for the overall survival of patients with Ov-HGSCs.

GO (MMS) ID	β	Cox-PH p-value	Log-rank p-value	CP	N1	N2	GO Name
Positive predictors							
GO:0000786	−0.67	4.4E-02	4.3E-05	1.4E-03	65	25	nucleosome
GO:0005765	−0.28	1.1E-02	3.7E-04	2.0E-03	232	135	lysosomal membrane
GO:0050900	−0.42	1.9E-03	3.8E-03	2.7E-03	110	92	leukocyte migration
GO:0007229	−0.44	3.0E-03	8.0E-03	4.9E-03	77	80	integrin-mediated signaling pathway
GO:0010923	−0.54	1.1E-02	3.1E-03	5.7E-03	51	49	negative regulation of phosphatase activity
GO:0007584	−0.63	3.4E-03	1.1E-02	6.1E-03	67	54	response to nutrient
GO:0007067	−0.36	4.5E-03	8.5E-03	6.2E-03	245	126	mitosis
GO:0007160	−0.41	8.2E-03	7.0E-03	7.6E-03	79	75	cell-matrix adhesion
GO:0000910	−0.58	2.0E-03	3.0E-02	7.8E-03	64	54	cytokinesis
GO:0006928	−0.49	1.0E-02	5.9E-03	7.8E-03	97	60	cellular component movement
GO:0001666	−0.31	1.4E-02	7.4E-03	1.0E-02	153	115	response to hypoxia
GO:0000922	−0.39	7.0E-03	1.6E-02	1.1E-02	86	77	spindle pole
GO:0031965	−0.33	1.7E-02	8.6E-03	1.2E-02	162	108	nuclear membrane
GO:0005813	−0.21	2.6E-02	7.4E-03	1.4E-02	353	165	centrosome
GO:0004674	−0.23	1.7E-02	1.3E-02	1.5E-02	374	217	protein serine/threonine kinase activity
GO:0070374	−0.36	3.1E-02	9.1E-03	1.7E-02	97	57	positive regulation of ERK1 and ERK2 cascade
GO:0044325	−0.37	9.1E-03	3.5E-02	1.8E-02	71	87	ion channel binding
GO:0055037	−0.63	1.5E-02	2.7E-02	2.0E-02	60	37	recycling endosome
GO:0004843	−0.46	1.8E-02	2.9E-02	2.3E-02	55	39	ubiquitin-specific protease activity
Negative predictors							
GO:0051436	0.55	2.5E-02	1.0E-03	5.1E-03	65	27	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
GO:0045666	0.15	4.0E-01	3.6E-03	3.8E-02	68	50	positive regulation of neuron differentiation
GO:0042393	0.17	3.6E-01	7.1E-03	5.0E-02	62	47	histone binding

β: the regression coefficients estimated by the Cox-PH model. **CP**: the composite p-value, which is the square root of the product of the Log-rank test p-value and the corresponding Cox-PH p-value. **N1**: the number of member genes covered by the corresponding MMS or GO term. **N2**: the number of mutated member genes present in at least one training sample. Note that a single gene can be covered by more than one GO term.

此为试读，需要完整PDF请访问：www.ertongbook.com