Tampere University

KALIFA MANJANG

Analysis of Prognostic Gene Expression Signatures of Breast and Prostate Cancer

Tampere University Dissertations 498

Tampere University Dissertations 498

KALIFA MANJANG

Analysis of Prognostic Gene Expression Signatures of Breast and Prostate Cancer

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Technology and Communication Sciences
of Tampere University,
for public discussion in the auditorium RG202
of the Rakennustalo building, Korkeakoulunkatu 5, Tampere,
on 19th November 2021, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Responsible supervisor and Custos	Assoc. Prof. Frank Emmert-Streib Tampere University Finland	
Supervisor	Prof. Olli Yli-Harja Tampere University Finland	
Pre-examiners	Prof. Dr. Andreas Holzinger Medical University Graz Austria	Assoc. Prof. Gang Hu Nankai University China
Opponent	Prof. Zlatko Trajanoski Medical University of Innsbruck Austria	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-2163-5 (print)
ISBN 978-952-03-2164-2 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-2164-2

PunaMusta Oy – Yliopistopaino
Joensuu 2021

Dedicated to the memory of my daughter, Taqiyah Manjang, whose memory will live in our hearts forever.

ACKNOWLEDGEMENTS

All praise to the Almighty for the numerous blessings, good health, strength, and endurance required to complete this dissertation.

I thank my supervisor, Prof. Frank Emmert-Streib, for providing me with a nurturing environment to grow as a researcher; his patience, honesty, encouragement, and openness kept me going during the difficult times. His counsel and advice have greatly influenced how I approach my work and life in general. I could knock on his door whenever I wanted, and he would gladly accept me. This work would not have been possible without his guidance, for which I will be eternally grateful. I would also like to thank Prof. Olli Yli-Harja, Dr. Shailesh Tripathi, Prof. Matthias Dehmer, and Prof. Galina Glazko for co-authoring the articles cited in this dissertation.

A warm thank you to the pre-examiners, Prof. Dr. Andreas Holzinger (Medical University Graz, Austria) and Assoc. Prof. Gang Hu (Nankai University, China), for taking the time to read this dissertation and provide valuable feedback, and to Prof. Zlatko Trajanoski, Medical University of Innsbruck, for agreeing to serve as the opponent during the public defense of this thesis.

A special thanks to Prof. Matthias Dehmer for fully funding my visit to the University of Upper Austria, Steyr campus. I had a great time during my stay and learned a lot from him. A special thanks also to Dr. Shailesh Tripathi for being such a great host during that time and for allowing me to ask questions and engage him in a discussion whenever I hit a brick wall. To my colleagues in the Predictive Society and Data Analytics Lab research group, Dr. Aliyu Musa, Samar Bashath, and Nadeesha Perera, thank you all for being such great and accomodating colleagues.

To my parents, thank you for a wonderful upbringing and for instilling in me the values of discipline, hard work, and never giving up. Your unwavering love has aided in the development of the person I am today. I'd also like to express my gratitude to my siblings for their support throughout this work. And I'd like to thank all of

my friends (you know who you are) for their invaluable support and encouragement over the years.

Finally, I want to thank my beautiful wife for her patience and unconditional love. Thank you for believing in me and encouraging me to pursue my dreams.

Tampere, August 2021
Kalifa Manjang

ABSTRACT

The advent of high-throughput sequencing and microarray technologies has improved our overall understanding of human health and diseases. Recent developments in 'omics' technologies (genomics, transcriptomics, proteomics, and metabolomics) have made it possible to measure tens of thousands of molecular quantities in parallel (in the same experiment). One specific objective of this thesis is to scrutinize a comparatively small subset of such measurements known as biomarkers or signatures relevant for predicting the course of a disease or patient's outcome.

In the context of cancer, the discovery of prognostic biomarkers for predicting cancer progression is an important problem for two reasons. First, such biomarkers may be used to treat patients in a clinical setting. Second, it is thought that investigating the biomarkers themselves would yield novel insights into disease mechanisms and the underlying molecular processes that trigger pathological behavior. The latter assumption is investigated in detail in this dissertation. Specifically, we study this problem for breast and prostate cancer by looking at a large number of previously reported prognostic signatures of breast and prostate cancer based on gene expression profiles. For this, we created a novel gene removal procedure (GRP) that purges all traces of biological meaning of the signatures genes and show that surrogate genes can be found among the remaining genes with better or equivalent prognostic prediction capabilities but distinct biological meaning as the published signature genes. As a result, our findings demonstrate that none of the examined signatures have a sensible biological meaning in terms of disease etiology and are merely black-box models allowing to make predictions of patient outcome but are not capable of offering causal explanations to improve disease understanding.

CONTENTS

List of Tables

ABBREVIATIONS

BM	Biomarker
BP	Biological Process
CC	Cellular Component
DAG	Directed Acyclic Graph
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
FPKM-UQ	Fragments Per Kilobase of transcript per Million mapped reads Upper Quartile
GDC	Genomic Data Commons
GO	Gene Ontology
GRP	Gene Removal Procedure
HER2	Human Epidermal Growth Factor Receptor 2
HGNC	HUGO Gene Nomenclature Committee
MF	Molecular Function
OS	Overall Survival
PC1	First Principal Component
PCa	Prostate Cancer
PFS	Progression-free survival
PG	Proliferation Genes
RGS	Random Gene Set
SG	Surrogate Genes
TCGA	The Cancer Genome Atlas

US	United States

ORIGINAL PUBLICATIONS

Publication I K. Manjang, S. Tripathi, O. Yli-Harja, M. Dehmer and F. Emmert-Streib. Graph-based exploitation of gene ontology using GOxploreR for scrutinizing biological significance. *Scientific reports* 10.1 (2020), 1–16. DOI: 10.1038/s41598-020-73326-3.

Publication II K. Manjang, S. Tripathi, O. Yli-Harja, M. Dehmer, G. Glazko and F. Emmert-Streib. Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. *Scientific Reports* 11.1 (2021), 1–18. DOI: 10.1038/s41598-020-79375-y.

Publication III K. Manjang, O. Yli-Harja, M. Dehmer and F. Emmert-Streib. Limitations of explainability of Prognostic biomarkers of prostate cancer. *Frontiers in Genetics* 12 (2021), 1095. DOI: 10.3389/fgene.2021.649429.

Author's contribution

This section describes the author's contribution in each of the original publications used in this dissertation. This thesis is based on the following articles, which are referred to in the text by the Roman numerals **I - III**.

Publication I — The author developed the R package used in Publication **I** called 'GOxploreR'; conducted the analysis and interpreted the results. In addition, he also participated in writing and editing the manuscript.

Publication II — The author performed the key role of wrangling the SWE dataset used in the publication, performed all the analysis as well as the interpretation of the results. He also contributed in writing and editing the manuscript.

Publication III — The author reviewed the literature for published prostate cancer prognostic signatures and gathered the gene expression dataset used in the analysis. He conducted the analysis and interpretation of the results, as well contributed in writing and editing the manuscript.

1 INTRODUCTION

In this chapter, the topics and motivation of this dissertation are discussed. In Section 1.1, a broad overview of the research is given. Section 1.2 address the research objectives, and finally, the structure of the dissertation is described in Section 1.3.

1.1 General background

Cancer is a complex disorder with many potential causes. Cancer tumors are heterogeneous, involving irregular growth of cells with the ability to metastasize or migrate to different body parts resulting in death. In the United States for example, 1,898,160 new cancer cases and 608,570 cancer mortality are estimated in 2021 [1]. Nevertheless, early diagnosis and general knowledge of cancer risk factors can help improve cancer survival rates immensely [2].

Tumor biopsy is the surgical removal of living tissue used as an invasive technique for cancer detection. But the side effects of biopsy, such as the risk of stimulating cancer progression, and metastasis have possible adverse effects on the patients [3]. Moreover, multiple biopsies to monitor disease progression and therapeutic response in patients diagnosed with cancer are almost impractical to perform. Besides, tumors found in vulnerable locations, such as the prostate, brain, and liver, require highly qualified medical practitioners to conduct surgical operations adding to the high expense attributed to tissue biopsy. As a consequence, the use of a safer, low-cost, and non-invasive or minimally-invasive procedure instead of tumor biopsy is a potentially useful approach.

The advent of high-throughput sequencing and microarray technologies have enhanced our comprehensive understanding of human health and diseases. Recent developments in 'omics' technologies (genomics, transcriptomics, proteomics, and metabolomics) have made it possible to measure tens of thousands of molecular quantities in parallel (in the same experiment). One specific objective of this study

is to determine a comparatively small subset of such measurements referred to as biomarkers or signatures relevant for predicting the course of a disease or patient's outcome. These subsets of measurements are referred to as features or feature selection in the statistics and machine learning field [4]. Biomarkers typically play an important role in early diagnosis, disease prevention, and the prediction and monitoring of treatment responses to different therapeutic interventions. Molecular biomarkers in particular are commonly known to be used in the study of human diseases [5]. The prognostic value of such predictions is quantitatively evaluated through a survival analysis, which supports a statistical test to be carried out to distinguish variations between different patient groups concerning 'time to event' information. The definition of 'event' is not limited to only death, but also relapse, disease progression, or organ failure. As a consequence of this phenomenon, prognostic studies are valuable for almost all patient-related medical investigations.

A significant number of molecular markers, especially prognostic markers, have been known in the literature capable of identifying cancer patients with good and bad prognoses. The main investigation of this dissertation is to confirm our hypothesis that prognostic signatures of breast and prostate cancer are lacking interpretability. In [6], it is stated that "A reliable set of predictive genes also will contribute to a better understanding of the biological mechanism of metastasis". This assertion is not confined to the issue above but is generally accepted to be true in the genomics and translational medicine community. Refuting this statement by showing that prognostic signatures of breast and prostate cancer are lacking a sensible biological meaning with respect to disease etiology is one of the main objectives of this dissertation.

1.2 Motivation and research objectives

In the previous section, cancer biomarkers are widely discussed and their general relevance in the clinical setting is indicated. This section highlights the research objectives of this thesis. The main objective of this dissertation is to demonstrate that the prognostic signatures of breast and prostate cancer in terms of disease etiology lack a reasonable biological definition.

1.2.1 Motivation and research objectives for Publication I

The gene removal technique introduced in Publication II [7] (and applied in Publication III) requires direct access to structural information of GO that enables the graph-theoretical properties of a DAG (directed acyclic graph) to be effectively exploited. Specifically, how to access the GO-term levels, categorizing GO-terms as jump nodes (JN), regular nodes (RN) and leaf nodes (LN), etc. So far there is no software available that provides such a functionality. For this reason, Publication I provides an R software package that contains functions for solving such problems.

1.2.2 Motivation and research objectives for Publication II

A large number of different prognostic signatures of breast cancer have been suggested. Specifically, Publication II [8] studies 48 such signatures and investigates their predictive capabilities and their biological meaning. For this analysis, we developed an approach that systematically purges all traces of biological meaning of signature genes, as measured by GO-terms (provided by the R package developed in Publication I).

1.2.3 Motivation and research objectives for Publication III

Due to the fact that the results obtained in Publication II are somewhat surprising we repeat a similar analysis for prostate cancer in Publication III [9]. This will reveal if those results hold only for breast cancer or if they do translate to other cancer types as well. Hence, in Publication III, we extend our analysis in Publication II to prostate cancer. This provides information about similarities and differences of the two cancer types. Furthermore, we discuss relations to the hallmarks of cancer.

1.3 Dissertation structure

This dissertation is outlined as follows. Chapter 1 introduces the general background and therapeutic relevance of biomarkers, Chapter 2 provides a review of the existing literature on breast and prostate cancer. The methodology of the analysis is presented in Chapter 3. Chapter 4 contains the summary of the results. A discussion

and concluding remarks are presented in Chapter 5.

2 REVIEW OF LITERATURE

2.1 Breast cancer

The breast is the tissue that overlies the pectoralis major muscle on the chest. The female breast consists of 15 to 20 sections, called lobes (in the lobes are smaller structures, called lobules), the fatty tissue, ducts, nipple, and the areola, which is the dark region around the nipple (see Figure 2.1). The primary role of the female breast is to produce milk for the young.

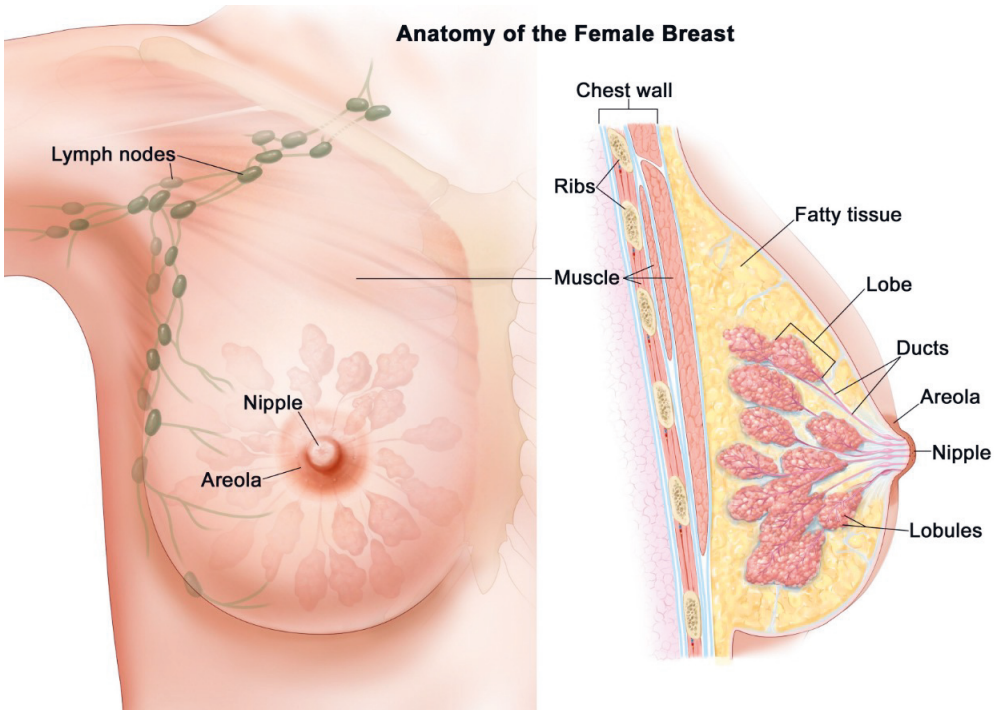


Figure 2.1 Anatomy of the Female Breast. Reprinted with permission from Terese Winslow LLC, Medical And Scientific Illustration [10].

Breast cancer (BC) originates from the abnormal growth of breast cells. Such cells divide faster than healthy cells do and begin to build up, forming lump or mass. BC can be benign or malignant. Benign breast tumors are abnormal growths, but they do not grow outside the breast. In contrast, malignant tumors are life-threatening and require immediate therapeutic interventions.

2.1.1 Epidemiology

Breast cancer is the commonest type of cancer in women and the leading cause of cancer-related death worldwide, with an estimated 2.3 million new cases diagnosed in 2020, meaning that the disease accounted for 1 in 4 cancer cases and 1 in 6 cancer deaths in women [11]. Women of various races, ethnicities, and geographical locations are prone to this cancer. As a consequence, the incidence, mortality, and survival rates vary considerably between different regions of the world, population structure, lifestyle, climate, and genetic factors (see Figure 2.2) [12]. In Finland, 4934 new BC cases were diagnosed in 2018, resulting in 873 causalities. The incidence rate is 176.75 per 100,000, and a five-year relative survival rate of 91% (Finnish Cancer Registry). By contrast, In the same year, 268,670 cases and 41,400 deaths was estimated in the US [13]. Breast cancer incidence is very low among men. Approximately, 1 in 1000 men will develop BC throughout their lifetime [14]. Interestingly, there is evidence of a decline in BC mortality in the developed countries since the early 2000s. This is possibly due to reduced use of hormone replacement therapy (HRT), early diagnosis by mammography tests, increased public awareness, and lifestyle improvements [15, 16].

2.1.2 Etiology

Gender and Age. Breast cancer majorly affects females. Therefore, gender is a big risk factor; apart from gender, age is one of the most significant risk factors. There is an elevation in the incidence of BC in older and middle-aged women. For this reason, women in this age category are more likely to develop the disease.

Race and Ethnicity. In the same way, race and ethnicity is also a risk factor. White non-Hispanic females have the highest incidence rates of BC, whereas BC mortality is 40% higher in Black American women relative to white women [1].

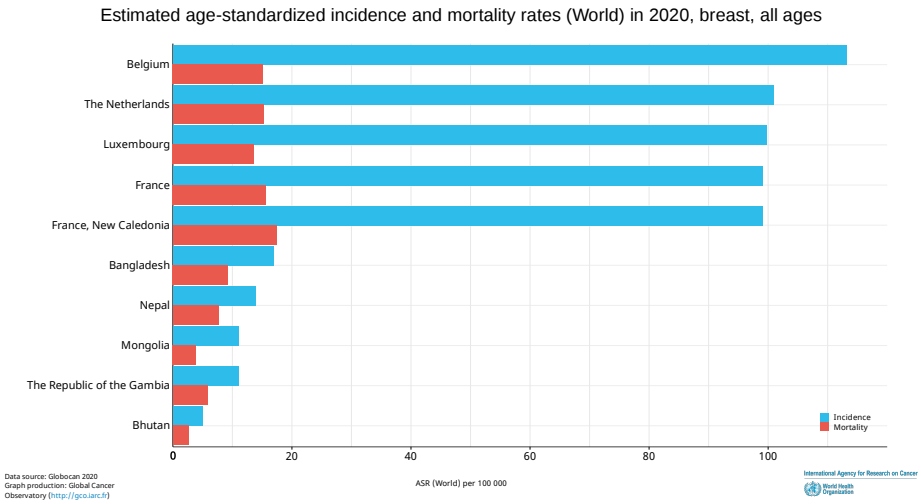


Figure 2.2 Estimated age-standardized incidence and mortality rates (World) in 2020, breast cancer across all ages. SOURCE: GLOBOCAN 2020 (IARC) [11].

Breast cancer history. A history of BC is also a risk factor, about 20 to 30 percent of women diagnosed, treated, and pronounced disease-free after initial local and regional treatment have a recurrence during follow-up [17].

Genetic predisposition. Researchers estimate that about a quarter of all breast cancer cases are family history related. Genetic predisposition accounts for approximately 5–10% of all breast cancers [18]. A number of inherited genes were identified which may increase the risk of breast cancer. The most common gene is the BReast Cancer gene 1 (BRCA1) and BReast Cancer gene 2 (BRCA2). Both of which greatly raise the risk of breast and ovarian cancer [19, 20].

Reproductive factors and lifestyle practices. Reproductive factors for women, such as never having a child, early menarche, late menopause, or late age at first pregnancy, may also increase the risk of breast cancer. Also, oral contraceptives and bad lifestyle choices, such as heavy alcohol consumption and excessive dietary fat intake, can also raise the risk of breast cancer. Alcohol intake can increase the number of hormones associated with estrogen in the bloodstream and can activate the estrogen receptor pathways [21].

2.1.3 Symptoms

The early symptoms and preclinical signs of breast cancer for most women vary from painless breast lump (palpable mass), discharge of the nipple other than breast milk, retraction of the nipple, changes in the appearance of either or both nipples, increase in size or alteration in the shape of the breast. For metastatic BC, the symptoms are determined by the body part the disease has progressed to and its stage. A lump may form anywhere other than the breast. Also, back or hip pain may develop. If the tumor spreads to the brain, neurological symptoms such as headache, memory loss, confusion, blurred or double vision, difficulty speaking, or seizure may arise.

2.1.4 Molecular subtypes

Breast cancer is a heterogeneous and complex disease that can be classified into different subtypes with distinct biological features, clinical behavior, and response to therapy [22]. Subtyping approaches include histopathology, molecular pathology, genetic analysis, and gene-expression profiling [23]. There are five main molecular subtypes of BC according to gene-expression profiling. These subtypes are luminal A, luminal B, HER2-overexpressing or HER2-enriched, basal-like, and normal breast-like tumors [24, 25]. **Luminal A** BC are estrogen receptor-positive (ER-positive), progesterone receptor-positive (PR-positive), HER2-negative, and low level of protein Ki-67. The tumor is low-grade, has the best prognosis, and a relatively low recurrence rates [26]. **Luminal B** is hormone-receptor positive i.e. ER-positive and (or) PR-positive. They may also be HER2-negative or HER2-positive with a high level of protein Ki-67. Luminal B accounts for roughly 10 percent of all BC and have a poorer prognosis compared to luminal A subtype [27]. **Triple negative/Basal-like** BC is hormone-receptor negative i.e. ER-negative and PR-negative and also, HER2 negative. Women with the BRCA1 (BReast Cancer gene 1) gene are more susceptible to this BC. The tumor tends to occur commonly in younger and black women [28, 29]. Basal-like tumors are more aggressive with a poorer prognosis compared to luminal A and luminal B tumors [30]. **HER2-enriched** BC are ER-negative and PR-negative and, HER2-positive. It can also be HER2-negative [31]. This cancer grows faster than luminal tumors and has the worse prognosis. **Normal-like** BC is hormone-receptor positive (ER-positive and (or) PR-positive) and, HER2-negative.

It also has low levels of Ki-67, which slows down the growth of the tumor. It is similar to the luminal A tumor but with a slightly worse prognosis (it still has a relatively good prognosis).

2.2 Prostate cancer

The prostate is a small walnut-shaped gland at the base of the urinary bladder (see Figure 2.3). The gland plays an important role in the male reproductive system. It helps produce semen, the milky fluid that nourishes and transports sperm. The three main prevalent forms of prostate disease are benign prostatic hyperplasia (BPH), prostate cancer, and prostatitis. Prostate cancer (PCa) is one of the commonest types of cancer. Some forms of the disease grow slowly and are localized to the prostate gland, and do not cause serious damage, while other types are more aggressive and can spread rapidly to other body parts leading to death. Early diagnosis of PCa is the best chance for successful treatment.

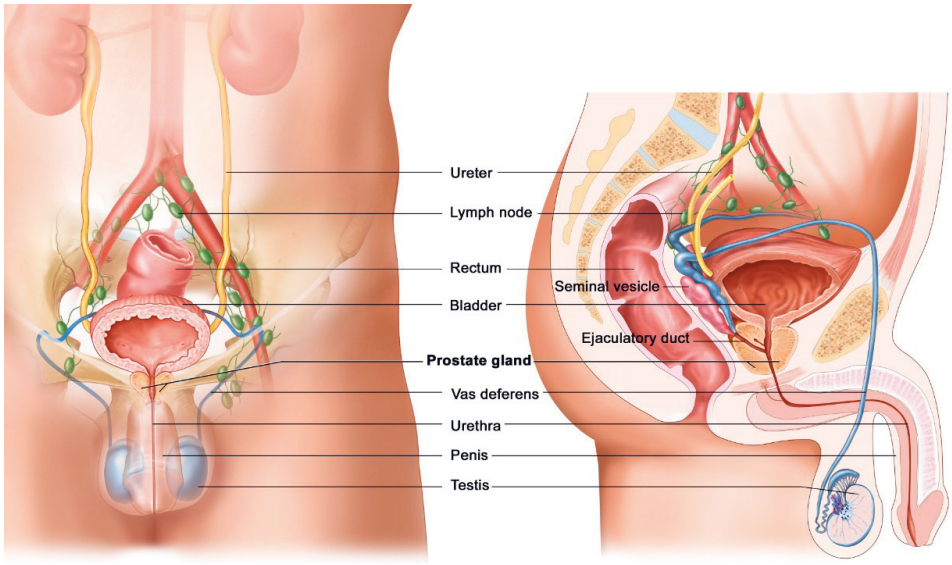


Figure 2.3 Anatomy of the Male Reproductive System. Reprinted with permission from Terese Winslow LLC, Medical And Scientific Illustration [10].

2.2.1 Epidemiology

Prostate cancer (PCa) is the second most common cancer in men and the fourth most common cancer worldwide [32]. The incidence rate of PCa increases with age. The average age of diagnosis is 66 years, and about 60 percent of all diagnosed cases occur in men over 65 years old [33]. Thus, it is common in countries with a higher proportion of older men. It is estimated that 191,930 new cases of PCa will be diagnosed in the US in the year 2020, resulting in about 33,330 deaths [34]. In 2020, 1,414,259 new cases of PCa are diagnosed globally, representing 7.3 percent of all cancers, 375,304 death globally was registered in the same year, accounting for 3.8 percent of all cancer deaths (GLOBOCAN 2020) [11]. The incidence and mortality rates of PCa are highly variable worldwide (see Figure 2.4). From the figure, the age-standardized rate (ASR) was highest in Northern Europe (83.4 per 100,000), and the highest mortality is in the Caribbean (27.9 per 100,000). The global variation of prostate cancer incidence can be linked to rampant prostate-specific antigen (PSA) testing in developed countries such as Europe and the US and less population-based testing in developing countries [35].

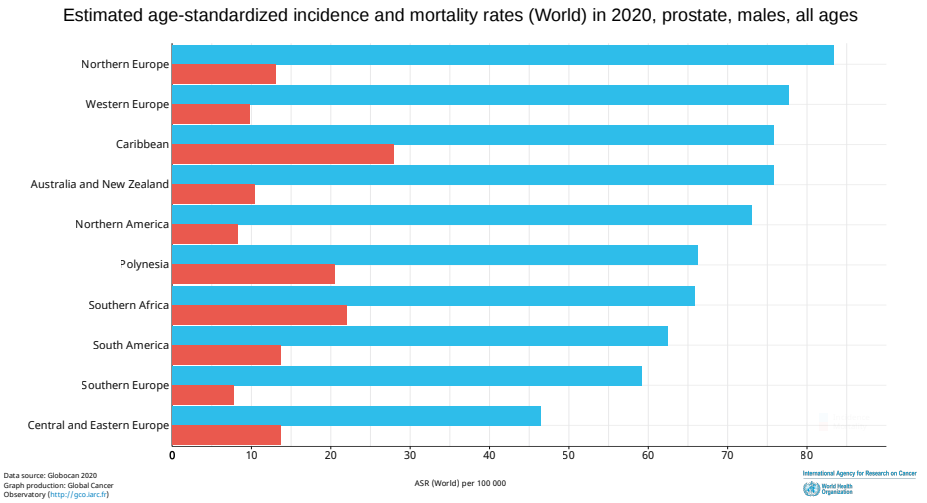


Figure 2.4 Estimated age-standardized incidence and mortality rates (World) in 2020, prostate cancer across all ages. SOURCE: GLOBOCAN 2020 (IARC) [11].

2.2.2 Etiology

Age. Prostate cancer is an age-dependent disease. In men younger than 50 years old, there is a moderate overall incidence of PCa [36]. About 60 percent of all diagnosed cases are in men of age 65 years and above, suggesting that the probability of men developing the disease quickly rises after they are 50 years old [37].

Race/Ethnicity. Prostate cancer incidence varies widely between various races, ethnic groups, and geographical locations. The lowest incidence rate of PCa is found in Asia with an incidence of 13.6 per 100,000, whereas the highest is in Northern America, an incidence of 73.0 per 100,000 (GLOBOCAN 2020) [11]. Black American males of African descent have the highest risk of prostate cancer globally and are more likely than other ethnic groups to develop the condition early in their lives [38].

Family history. Men with a history of PCa in their family are higher at risk of PCa than men with no family history of PCa [39]. This risk is even higher for men with relatives diagnosed at a younger age [37]. Also, genetic predisposition such as having a strong family history of genes such as BRCA1 and especially BRCA2 can increase the risk.

Diet. Dietary factors can play an essential role in PCa growth. Consumption of saturated animal fat, red meat, calcium, milk, and dairy products have been related to an elevated risk of PCa [40, 41, 42].

2.2.3 Symptoms

Although certain men are asymptomatic and hence, do not experience any signs or symptoms, others in the early stages of the disease may experience discomfort. The nature of the symptoms largely depend on the location of the tumor in the prostate and whether or not it is indolent. The symptoms of PCa vary from frequent urination, increased desire to urinate, especially at night, burning or pain when urinating, slow or delayed urine flow, blood in the urine or semen. If the disease progresses past the prostate, there may be signs such as hip, back, and shoulder pain, as well as excessive weight loss and exhaustion [43].

2.3 Definition of prognostic biomarkers

2.3.1 What are biomarkers?

According to the National Cancer Institute, a biomarker is 'a biological molecule found in the blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease'. In other words, biomarkers provide measures of biological induced changes in the body that is indicative of disease progression or other health concerns. A biomarker can be molecular, physiological, or biochemical. Also, they can either be a single entity, e.g. the mutation of a single gene, BRCA1/2 in BC, or a collection of those entities called biomarkers or a set thereof. Importantly, if one has a set of biomarkers, the entities in such a set are typically of a common type, i.e. all relating to the levels of proteins or the expressions of genes or genetic mutations. Biomarkers can mainly be classified into three types: diagnostic, predictive, and prognostic (see Figure 2.5).

Diagnostic biomarkers are used to evaluate an unknown condition and to assess the progression of the disorder and/or the efficacy of a treatment intervention. On the other hand, prognostic biomarker according to [44] are characterized as a marker 'used to identify and classify a patient according to their level of risk of an outcome of interest in the absence of treatment', and predictive markers 'are used to identify and classify patients to predict an outcome of interest in response to a particular treatment'. Similarly, [45] also described prognostic biomarkers as 'a clinical or biological characteristic that provides information on the likely patient health outcome (e.g. disease recurrence) irrespective of the treatment'. In the same way, predictive biomarker 'indicates the likely benefit to the patient from the treatment'. Based on the aforementioned definitions, it is clear that biomarkers can be used in defining patient outcomes. Ideally, we treated such patient groups as a two-class pattern classification (alive vs deceased). The emphasis of this work is on prognostic biomarkers.

Prognostic biomarkers are capable of binary grouping of patients and the variation in progression of these patients in these two groups due to the disorder can be detected through survival analysis. Typically, such groups may demonstrate a varying survival time to event, considering that the meaning of 'event' and 'survival time' are contextual. Specifically, an 'event' could refer to death, progression, or relapse. The survival times are called overall survival (OS - death), progression-free survival

(PFS - worsening of a disorder), relapse-free survival (RFS - recurrence of disorder), or disease-free survival (DFS - duration of disease-free status) depending on the nature of the event. Accordingly, statistical approaches have been developed to classify the disparities in survival times of the different patient groups, which are ultimately accomplished by comparing the Kaplan Meier curves of the corresponding two patient groups.

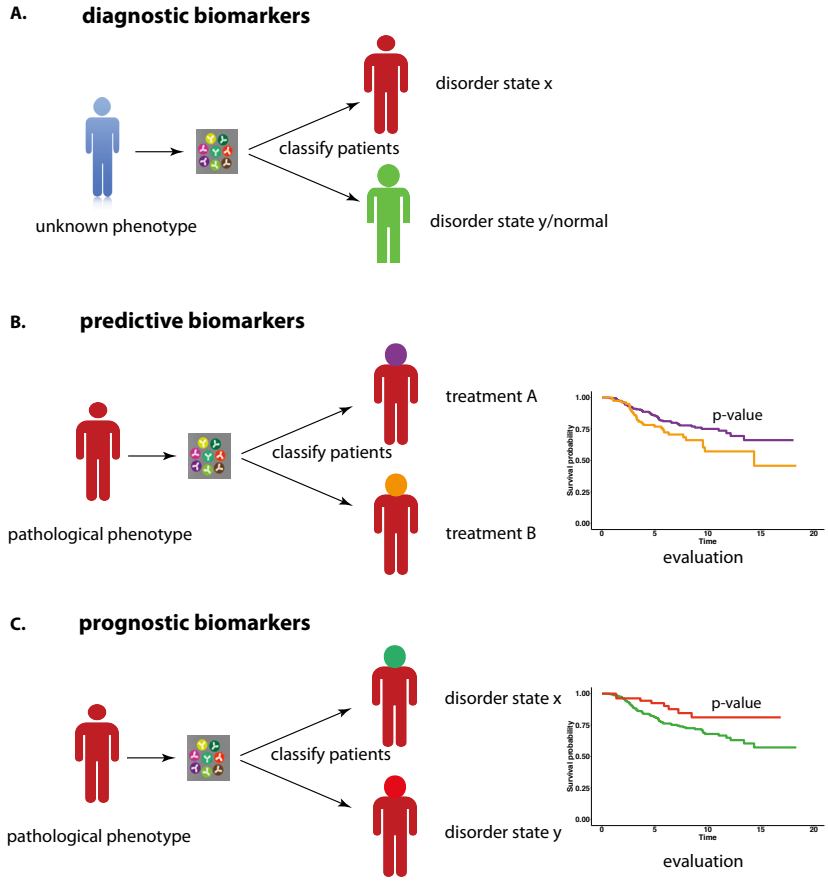


Figure 2.5 Overview of the three types of biomarkers. **A**: Diagnostic biomarkers. **B**: Predictive biomarkers. **C**: Prognostic biomarkers.

One can infer from this discussion that to study prognostic biomarkers, certain things are needed: (i) at least two separate patient groups, whereas (ii) each group represents a particular state of a disorder. For instance, triple-negative vs non-triple negative breast cancers [46]. The statement is no suggestion that prognostic biomarkers are only useful given the specific states of the disorder. Interestingly, prognostic

biomarkers are also useful where the specific states of the disease are uncertain. Only its heterogeneity is needed.

2.3.2 Identifying prognostic biomarkers

A Pubmed search at the time using keywords 'prognostic biomarkers' indicates that more than 85,000 papers have investigated prognostic markers. This is a large number of publications. However, the central underlying design of the biomarkers investigated in all these papers can be outlined by a general procedure [47, 48]. This fundamental procedure is shown in Figure 2.6.

1. Generation of gene expression data
2. Preprocessing of the data
3. Selection of biomarkers
4. Categorization of patient samples
5. Assessment of the biomarkers

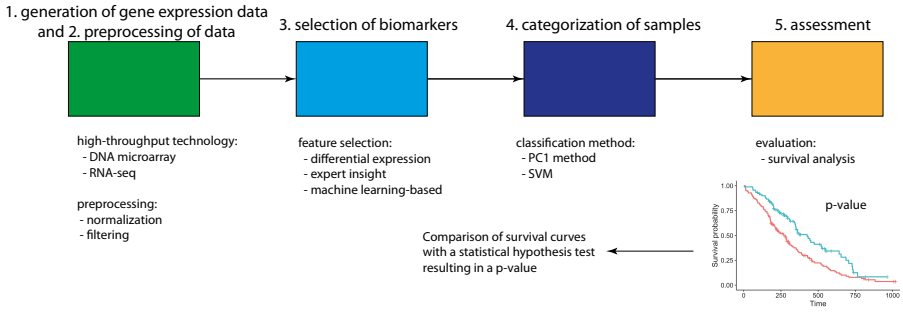


Figure 2.6 General technique used in studies to create prognostic biomarkers.

The procedure mentioned above is consistent with the general approach used when prognostic signatures are identified or studied. All studies adopt these steps, which vary only in terms of the methodology used. Studies have, for example, used various methods to select genes as prospective biomarkers. A widely used approach involved the identification of differentially expressed genes, an example is from a study by [49]. Firstly, the authors studied the TCGA data and 60 genes were identified that were differentially expressed between favorable and unfavorable prognosis and maybe a likely biomarker for the prognosis of prostate cancer. The subsequent

analysis also included the discovery of differentially expressed genes using the lymph node-negative LAPC samples from Russian patients. They discovered 12 additional differentially expressed genes as potential prostate cancer markers. Similarly, other studies by [50, 51, 52, 53, 54], and [55] all use differentially expressed genes to identify potential prognostic biomarkers for prostate cancer, breast cancer, gastric cancer, and pancreatic carcinoma. Another approach, based on machine learning, feature selection, in general, is proposed by [56, 57]. Another example applies to the categorization of the patients (samples from patients) using various classification methods, such as SVM has been investigated by [58, 59, 60]. The PC1 method can also be used for this purpose. For instance, in [61] the PC1 of the signatures are computed and the cohorts are splits according to the median of PC1.

Intriguingly, only one approach namely a survival analysis is employed for determining the prognostic significance of biomarkers [62]. Specifically, a comparison of various Kaplan Meier survival curves is made using a statistical hypothesis test that measures considerable heterogeneity in these curves. Generally, only two survival curves are compared, referring to two classes of patients with distinctive prognostic behavior, although it is possible to expand them to more groups.

3 MATERIALS AND METHODS

The materials and methods used in Publication I - III are presented in this chapter. In Section 3.1, the gene expression data used in Publication II and III is discussed. Section 3.2 introduce the breast and prostate cancer prognostic biomarkers that are used in the study. Section 3.3 explains how the GO-DAG used for retrieving the GO-terms hierarchy levels in Publication I is generated. Finally, the methods are discussed in Section 3.4.

3.1 Gene expression data (II, III)

Two types of publicly accessible gene expression data are used for the analysis - DNA microarray and RNA-seq data. In Publication II, we used both types of gene expression data. Publication 3 used only RNA-seq data during the analysis.

3.1.1 DNA microarray data

The first set of gene expression data set used is the NKI data set, accessible from [61]. The data set contains breast cancer samples from the Netherlands Cancer Institute (a.k.a NKI) cohort. The total number of unique samples is 295. Each gene expression profile had gene expression data for 13,108 genes, and each sample corresponds to a single patient. The patients have stage **I** or **II** breast cancer. The data set is supplemented by survival data, and the development of metastases indicates an 'event' for survival analysis. This data set was used in Publication II.

3.1.2 RNA-seq data

The second set of gene expression data used in Publication II is a breast cancer data set from the Gene Expression Omnibus with accession number **GSE96058** [63]. We

refer to this data as the SWE data. We opted for this data set due to its availability and the large number of samples it constitutes. In comparison to the NKI breast cancer cohort, it is a more recent data set. The data were FPKM normalized and log-transformed. It has the gene expression profiles of 30,865 genes and samples of subtypes: Basal, Her2, LumA, LumB, and Normal. Each subtype contains 360, 348, 1709, 767, and 225 samples respectively. The samples of normal subtypes are excluded from the data set. Also, genes without an associated Entrez gene ID are equally omitted. All genes had expression data available across all the samples. The overall survival endpoints are used during survival analysis.

The two RNA-seq data sets - HTSeq-FPKM and HTSeq-FPKM-UQ, used in Publication III contains PCa samples from The Cancer Genome Atlas Prostate Adenocarcinoma (TCGA-PRAD) project. The data was obtained from the UCSC Xena GDC data hub . The two data sets were FPKM and FPKM-UQ normalized and log-transformed respectively. It contains 551 samples, of which 498 are primary solid tumors, 52 are solid tissue normal, and just one metastatic sample. The metastatic and solid tissue normal samples are excluded from the final gene expression data set. Genes with no expression data across all samples are omitted. After filtering, the final gene expression profiles had gene expression data for 16,428 genes for HTSeq-FPKM and 15,165 genes for the HTSeq-FPKM-UQ data set. These data sets are referred to as GDC cohort A and B. For survival analysis, we used the progression-free survival endpoints. The patient survival information is provided by [64].

3.2 Biomarkers (II, III)

The second source of data used in the analysis is the prognostic biomarkers of breast and prostate cancer. The 48 prognostic biomarkers used in Publication II are compiled by [61] from 46 published studies through a literature search. Together the 48 biomarkers contain 8106 genes. In Publication III, the prostate cancer gene signatures reported are identified from Pubmed using keywords 'prognostic', 'biomarkers', 'signatures' and 'prostate cancer'. The search resulted in the compilation of 32 prognostic signatures from 31 studies. The biomarkers are reported using the HGNC gene nomenclature. As a result, the corresponding Entrez gene IDs of each biomarker set are derived. Table 3.1 and 3.2 contains the respective breast and prostate cancer signatures that we have used.

3.2.1 Published breast cancer studies analysed in Publication II

Acronym for a study	Number of genes	Cancer type	Reference
ABBA	111	Breast cancer	[65]
ADORNO	2	Breast cancer	[66]
BEN-PORATH-EXP1	367	Breast cancer	[67]
BEN-PORATH-PRC2	631	Breast cancer	[67]
BUESS	30	Breast cancer	[68]
BUFFA	3	Multiple cancers	[69]
CARTER	70	Multiple cancers	[70]
CHANG	355	Multiple cancer	[71]
CHI	136	Multiple cancer	[72]
CRAWFORD	377	Breast cancer	[73]
DAI	35	Breast cancer	[74]
GLINSKY	11	Multiple cancer	[75]
HALLSTROM	78	Multiple cancer	[76]
HE	6	Breast cancer	[77]
HU	13	Breast cancer	[78]
HUA	1345	Breast cancer	[79]
IVSHINA	17	Breast cancer	[80]
KOK	179	Breast cancer	[81]
KORKOLA	21	Breast cancer	[82]
LIU	167	Multiple cancer	[83]
MA	30	Breast cancer	[84]
MILLER	18	Breast cancer	[85]
MORI	156	Multiple cancer	[86]
PAIK	16	Breast cancer	[87]
PAWITAN	46	Breast cancer	[88]
PEI	2	Breast cancer	[89]
RAMASWAMY	16	Multiple cancer	[90]
REUTER	714	Breast cancer	[91]
RHODES	67	Multiple cancer	[92]
SAAL	162	Multiple cancer	[93]
SHIPITSIN	56	Breast cancer	[94]
SORLIE	15	Breast cancer	[95]
SOTIRIOU-93	343	Breast cancer	[96]
SOTIRIOU-GGI	90	Breast cancer	[97]
META-PCNA	129	Multiple cancer	[98]
TAUBE	242	Breast cancer	[99]
TAVAZOIE	6	Breast cancer	[100]
VALASTYAN	6	Breast cancer	[101]
VANTVEER	60	Breast cancer	[102]
WANG-76	69	Breast cancer	[103]
WANG-ALK5T204D	239	Breast cancer	[104]
WELM	3	Breast cancer	[105]
WEST	468	Breast cancer	[106]
WHITFIELD	587	Breast cancer	[107]
WONG-ESC	335	Breast cancer	[108]
WONG-MITOCHON	217	Breast cancer	[109]
WONG-PROTEAS	46	Breast cancer	[109]
YU	14	Multiple cancer	[110]

Table 3.1 A summary of the published prognostic signatures for Breast cancer used in Publication II. These signatures genes are derived by [61].

3.2.2 Published prostate cancer studies analysed in Publication III

Acronym for a study	Number of genes*	Cancer type	Reference
AGELL	12	Prostate cancer	[111]
BIBIKOVA	16	Prostate cancer	[112]
BISMAR	12	Prostate cancer	[113]
CHEN	4	Prostate cancer	[114]
CHEN_CC	7	Prostate cancer	[115]
CHEVILLE	2	Prostate cancer	[116]
CHU	8	Prostate cancer	[117]
CUZICK	31	Prostate cancer	[118]
GLINSKY	11	Multiple cancers	[75]
IRSHAD	19	Prostate cancer	[119]
IRSHAD_1	3	Prostate cancer	[119]
LARKIN	7	Prostate cancer	[120]
LI	6	Prostate cancer	[121]
LIU	167	Multiple cancers	[83]
LONG	12	Prostate cancer	[122]
NAKAGAWA	17	Prostate cancer	[123]
PENNEY	157	Prostate cancer	[124]
RAMASWAMY	16	Multiple cancers	[90]
REDDY	16	Prostate cancer	[125]
ROSS-ADAMS	100	Prostate cancer	[126]
ROSS	6	Prostate cancer	[127]
SAAL	162	Multiple cancers	[93]
SHARMA	15	Prostate cancer	[128]
SINGH	5	Prostate cancer	[129]
SONG	15	Prostate cancer	[130]
STEPHENSON	10	Prostate cancer	[131]
TALANTOV	3	Prostate cancer	[132]
TANDEFELT	36	Prostate cancer	[133]
TRUE	86	Prostate cancer	[134]
WANG	43	Prostate cancer	[135]
WU	29	Prostate cancer	[136]
YU	14	Multiple cancers	[110]

Table 3.2 A summary of the published prognostic signatures for prostate cancer used in Publication III. Number of gene* corresponds to the number of genes following HGNC gene conversion to Entrez gene IDs [9].

3.3 Gene Ontology (I-III)

The gene ontology (GO) is an important database mostly used to provide biological interpretations for the analysis of genes or gene set from biological, medical, and clinical problems. Originally, GO supported only three model organisms but has since been expanded to more than 3200. GO explains our understanding of the biological

domain in three distinct aspects of gene function, namely biological process (BP), molecular function (MF), and cellular component (CC) together with over 45,000 terms and 130,000 relations. The majority of the information, however, is based on ten model organisms (human, mouse, rat, zebrafish, drosophila, caenorhabditis elegans, dictyostelium discoideum, saccharomyces cerevisiae, schizosaccharomyces pombe, arabidopsis thaliana, and escherichia coli) [137]. Moreover, GO includes annotations by relating particular gene products to GO-terms. This facilitates the connection between genes and GO-terms for the derivation of knowledge unique to the organism.

3.3.1 Exploring the GO-DAG

Before the GOxploreR package, existing tools do not provide tailor-made functions to explicitly obtained the GO-DAG for the three sub-ontologies - BP, MF, and CC, the hierarchy level of a GO-term, creating organism-specific GO-DAGs, or providing means of visualizing the entire GO tree. Due to this, structural analysis of GO is extremely difficult for beginners or tedious for the experienced user. To this end, special functions are created in the resulting R package of Publication I for this purpose. Additionally, voiding the genes in a biomarker of all biological meaning requires selecting genes in the gene pool in a defined or constraint manner (Publication II and III). This approach demands, first and foremost, that:

1 All genes from the respective biomarker be removed.

2 Secondly, genes that belong to the same biological processes as the genes in the BM are omitted. According to the GO database [138], the biological processes are hierarchically ordered. Thus, genes of biological processes on the same hierarchy level are successively removed.

The GO-DAG needs to be deduced so that the structural information of GO can be investigated for this purpose. Taking into account that a child term of a parent node does not necessarily have to be on the next hierarchy level after the parent and can be further down the DAG (jump node). The terms jump nodes (JNs), regular nodes (RNs), and leaf nodes (LNs) are used to describe GO-terms with these distinct characteristics. These features demonstrate how one can create the GO-DAG from this information for every domain, i.e. biological process, molecular function, and

cellular component. See Publication I for a detailed description of the algorithm used to create the GO-DAGs.

3.4 Statistical analysis (II-III)

An outcome association method was created for determining the prognostic importance of random gene sets (RGS). This procedure consists of three key steps. First, selection/construction of a random gene set (see the gene removal procedure described in Publication II and III), categorization of patient samples, and, finally a survival analysis. Also, multiple testing correction after survival analysis was conducted using conservative Bonferroni correction. All statistical analyses were conducted using the R programming language [139].

3.4.1 Selection/construction of the gene set

Random gene sets (RGS) are chosen or constructed from a gene pool by first eliminating both the biomarker signatures and the genes that belong to the same biological processes as the genes in the BM signatures. The proposed method for the gene removal procedure used is described explicitly in Publication II and III.

3.4.2 Unsupervised classification

An unsupervised classification method by the PC1 method was used to stratified the patient samples into two classes (low and high risk). The PC1 approach is based on the principal component analysis which is a dimensional reduction technique. This method aims to transform a large data set into a smaller one having a lower-dimensional representation. The R function 'prcomp' was used to obtain the first principal component (PC1) of the signature. Based on the median of PC1, the patients are classified into two groups.

3.4.3 Survival analysis

To assess the prognostic importance of RGS, survival analysis is carried out. More precisely, a Kaplan Meier estimate of survival curves is performed and compared

with the Mantel-Haenszel test [140]. Each comparison is thus distinguished by a p-value that comes from such a hypotheses test. The R package 'survival' was used for survival analysis. P-values less than 0.05 were considered to be statistically significant. The categorization of the patient samples for survival analysis was achieved by the PC1 method as described earlier.

4 SUMMARY OF THE RESULTS

This chapter summarizes the findings of Publication I-III [7, 8, 9]. First, a brief description of some of the functionalities of the R package (GOxploreR) which resulted from the analysis of Publication I is described. Next, a summary of the results from Publication II and III is given. Finally, the contribution of this dissertation is discussed.

4.1 GOxploreR: An R package for the structural exploration of GO (I)

In Study I [7], GOxploreR, an R package was developed to facilitate the structural exploration of GO. The package provides support for ten species corresponding to the main organisms within the GO database. The package includes features for mapping gene or gene list to GO-terms and their associated hierarchy levels. If a list of GO-terms is generated, GOxploreR provides special functions that can help to obtain the corresponding hierarchy levels of these GO-terms. As already noted, the child of a GO-term can jump levels. The function 'GOTERMXX2ChildLevel' is useful for giving the child terms of a GO-term and its respective hierarchy levels. The 'XX' in the name can be substituted for BP, MF, or CC. GO offers different unique DAG for several species because each organism has a specific number of genes. For instance, human has approximately 20,000 - 25,000 genes and mouse has 30,000. From these genes, one can derive only a subset of all GO-terms that are connected to a specific organism. GOxploreR contains 11 DAG, one main DAG, and ten DAGs for each of the organisms that the package supports. The number of genes, total number of hierarchy levels, and GO-terms of BP for each DAG varies. In GOxploreR v1.1.0, the human GO-DAG of BP contains 19,155 genes, 19 hierarchy levels, and 12,436 GO-terms of BP. Similarly, E.coli GO-DAG of BP contains 3,449 genes, 15 hierar-

chy levels, and 1,491 GO-terms of BP. A detailed overview of the other supported organism is provided in Publication I.

For enriched GO-terms analysis, one needs to restrict such analysis to more detailed GO-terms that are placed at a higher hierarchy level. The identification of enriched GO-terms for a list of genes is perhaps the most important function of GO. It is not unusual to discover a vast number of such GO-terms, making a focused discussion very challenging to address. Nonetheless, the GO-DAG contains details that can be utilized for exploratory analysis of such a list. In particular, the hierarchy levels of GO-terms can be used. Although a GO level is not an absolute biological indicator, the information provided remains valuable [141]. For this purpose, GO-terms located on a specific hierarchy level can be obtained. Enabling, for example, a basic ordering of these GO-terms to supplement an enrichment analysis.

4.1.1 Visualization capabilities of GOxploreR

The most challenging aspect of visualizing a GO-DAG is the vast number of GO-terms such a DAG contains, making the visualization task infeasible. The reduced GO-DAG is introduced to tackle this problem. The underlying idea of such a GO-DAG is that the visualization problem is approached by mapping GO-terms into three-node categories, namely JN, RN, and LN. For example, the GO-terms of BP containing 12,436 nodes can be visualized using only 52 node categories (see Figure 4.1). Only category nodes containing at least one GO-term are shown, enabling a system-wide view of all human GO-terms of BP. In the same way, there is another function that visualizes only sub-GO-DAGs unique to an organism, comprised of only GO-terms of interest. These overall functionalities make GOxploreR versatile for the structural exploration of GO. A summary of the main functionalities provided by the package is as follows:

1. direct access to structural features of GO
2. structure-based ranking of GO-terms
3. mapping to a reduced GO-DAG
4. prioritizing of GO-terms

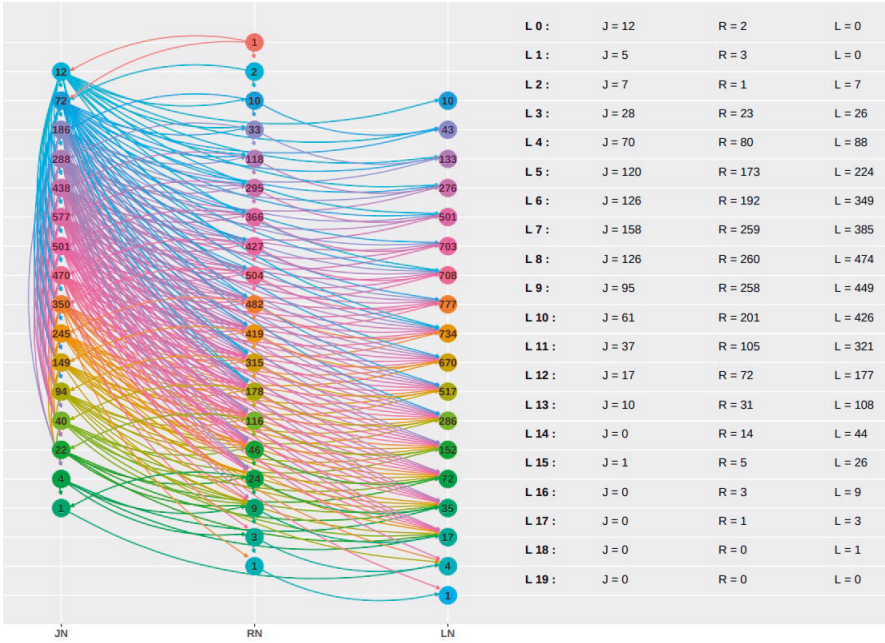


Figure 4.1 A reduced GO-DAG of BP for human. The whole GO-DAG contains 52 nodes, i.e. RN, JN, LN and it summarizes all the 12,436 GO-terms of BP available for this organism.

4.1.2 GOxploreR accessibility

The GOxploreR package is freely available from CRAN (Comprehensive R Archive Network). A thorough introduction to the package functions can be accessed from the package vignette (`https://cran.r-project.org/web/packages/GOxploreR/index.html`). The package can be installed by typing 'install.packages("GOxploreR")' on the R console.

4.2 Prognostic signatures of Breast cancer (II)

In study II [8], we systematically showed that the prognostic signatures of breast cancer outcome do not have a sensible biological meaning with regard to disease etiology. The study achieved its results using two BC gene expression data sets and a set of biomarker genes from 46 studies. An exploratory analysis of the genes in the biomarkers has shown that the size of biomarkers ranges significantly from one study to another. A pairwise study of the BM signatures has also shown that none of

the biomarkers are unique. Interestingly, all of them shared genes in common with at least 2 other biomarkers. On the other hand, the overlap of the GO-terms among the signatures was also explored. Importantly, the signatures shared at least some GO-terms with every other signature proving that all the signatures have a non-zero overlap in their biological meaning as measured by GO-terms.

The prediction capabilities of RGS are investigated by randomly sampling genes from the gene pool and conducting a survival analysis. The size of the respectively published biomarkers determines how many genes are sampled per random signature. The sampling is repeated 1000 times for each study, meaning we have analyzed $48,000$ RGS that are produced in this manner. The outcomes of the study are summarized in three ways. First, the p-values derived from the survival analysis are reported without any multiple testing corrections performed (uncorrected). In the study by [61], the same analysis was done i.e. no multiple testing corrections were applied to the p-values obtained. Next, we repeated the analysis by applying a conservative Bonferroni correction to the p-values (corrected) and finally, the proliferation genes are removed from the gene pool, and the p-values from the survival analysis are corrected using conservative Bonferroni correction (corrected without proliferation genes (PG)).

The results of the analysis summarized in Figure 4.2 are for the SWE data using gene removal procedure 2 (GRP 2) discussed in Publication II, where in addition to BM signatures, genes from the same biological process as the genes in the BM signatures are eliminated. The results are for the last hierarchy level, where all biological meaning from the RGS has been removed. In the figure, the red/green points reflect the results of the original BM signatures, with dark red/dark green suggesting non-significant values and light colors indicating significant results. The violet distributions correspond to the outcomes of random signatures and the shaded green bars correspond to the lower 3rd percentile of these distributions. Also, the horizontal black lines represent the median values of the distribution of random signatures and the long horizontal blue line corresponds to the significance level of $\alpha = 0.05$. The p-values are on a logarithmic scale (i.e. $\log_{10}$). From the results obtained, it is evident that not all BM signatures (big points) lead to significant outcomes. An explanation for this is that a different validation data than used by the 48 biomarkers studies is used. Despite this, still, in all instances (see Figure 4.2) 34 BM signatures are significant. As a consequence, 14 signatures do not show prognostic significance for

the independent validation data and therefore, lack robustness (For the NKI data, 39 published BM signatures are significant and 14 are non-significant). RGS with similar or better prognostic prediction capabilities as the published BM signatures are referred to as surrogate gene sets (SG). The SG are not assigned (biological) meaning or role, yet have equivalent or better prediction capabilities as the reported BM signatures as indicated by the green shaded bar in Figure 4.2, but entirely different biological interpretation. The NKI BC gene expression data indicate the same conclusion, implying that our findings are robust.

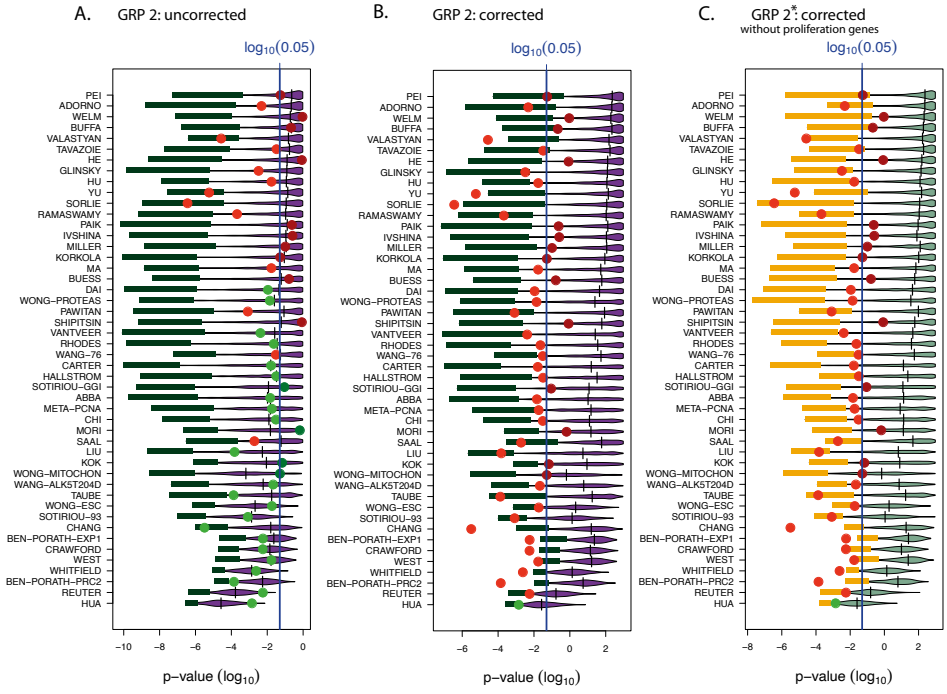


Figure 4.2 Results for gene removal procedure 2 and the SWE data. The patients samples contain LUM A and LUM B BC subtypes. The results in **A** are for uncorrected p-values and **B** for Bonferroni corrected p-values. In **C**, the proliferation genes are removed and the p-values are corrected [8].

4.3 Prognostic signatures of Prostate cancer (III)

In Study III [9], we investigated PCa prognostic signatures concerning their lack of biological meaning. The gene removal method described in Publication II was

used to systematically extract all biological associations between the reported BM signatures and genes in the gene pool from which random genes are chosen. The published signatures used for the analysis exhibited different characteristics than the published BC signatures discussed in Publication II. In particular, the size of the reported PCa signatures are smaller, all less than 200 genes (Table 3.2). A pairwise similarity study of the GO-terms in the published PCa signatures suggests that the signatures are more similar at the GO-term level than at the gene level.

Likewise, the outcomes of the investigation was presented in three parts. First, the results of the uncorrected p-values are presented, followed by the correction of these p-values by conservative Bonferroni correction. Finally, we show the outcomes of the optional step of the gene removal procedure discussed in Publication II, i.e. the PG are removed and the p-values are corrected. The results are shown in Figure 4.3 for the GDC cohort A data. The light/dark red bold points are the outcome of the previously reported PCa signatures and the new validation data. The blue-colored distributions are the outcomes of RGS, with the shaded cyan bars representing the lower third percentile of the distributions and the bold black points are the median of these distributions. The blue vertical line corresponds to a significant level of $\alpha = 0.05$. The results indicate that all published signatures (red points) do not lead to a significant outcome. Specifically, 24 and 22 BM out of the 32 published BM signatures are significant for the GDC cohort A and B data respectively, implying that the remaining signatures lack robustness for the independent validation data set.

In all three cases i.e, uncorrected p-values, corrected p-values, PG are removed and the p-values are corrected, we found many RGS with similar or better prognostic prediction capabilities as the PCa published signatures proving our hypothesis right that the published PCa signatures examined lack a sensible biological meaning because we were able to find RGS with no assigned biological meaning that provided the same outcome. We repeated the same analysis for the GDC cohort B data. The outcomes of which confirmed our analysis that RGS can always be found that yield significant outcomes.

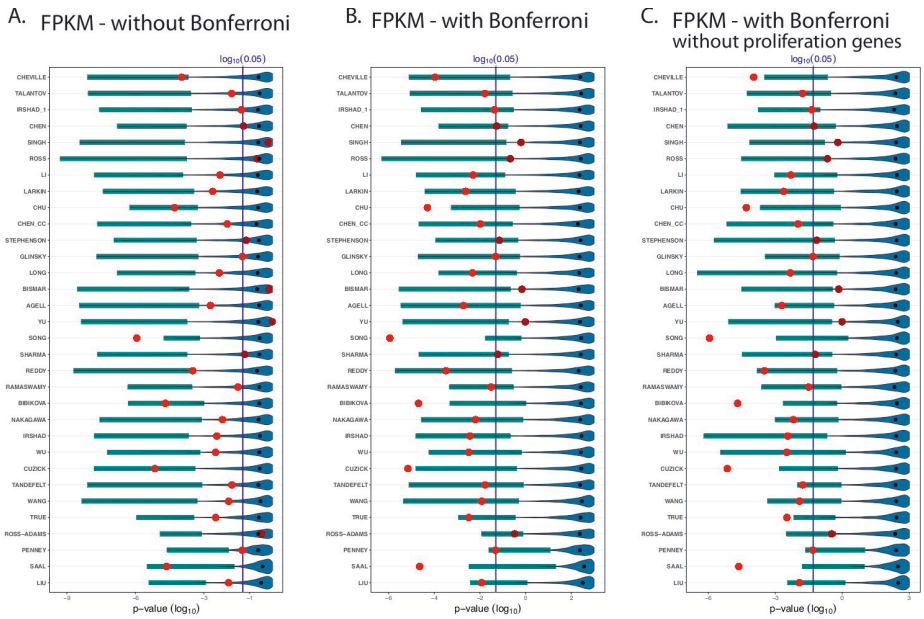


Figure 4.3 Results for the prediction capabilities of random gene sets. **A**: Results for uncorrected p-values. **B**: Bonferroni corrected p-values. **C**: Proliferation genes are removed and the p-values are Bonferroni corrected [9].

5 DISCUSSION AND CONCLUSION

This chapter provides a discussion and concluding remarks of this dissertation. The chapter begins with a section discussing some of the functionalities that our R package offers. Furthermore, the problems in the interpretation of prognostic biomarkers of breast and prostate cancer are highlighted. Finally, concluding remarks are given.

5.1 Discussion

5.1.1 GOxploreR for scrutinizing biological significance (I)

The GO knowledge-base has become increasingly important in recent years for providing annotations as the prevalence of data generated by high-throughput techniques has increased. Due to this, tailor-made tools that can enable efficient analysis of GO are needed. Deriving structural information of GO can be tedious even for the experienced user and extremely challenging for the novice user. For this purpose, our tool [7] provides functionalities for easy access to structural details of GO with respect to the underlying DAG (directed acyclic graph). In comparison to other packages, our package GOxploreR is significantly different. Specifically, one of the most unique features of GOxploreR is its ability to map a GO-DAG to a *reduced GO-DAG*. Also, the visualization of reduced GO-DAGs for the three sub-ontologies can easily be obtained. The reduced GO-DAG is useful for providing an overview of the biological information processing of the entire ontology. To our knowledge, no software tool for analyzing GO provides such capabilities. In addition, GOxploreR includes a prioritizing algorithm for removing GO-terms that capture redundant information from a list of GO-terms. The algorithm utilizes the fact that higher GO-terms capture more specific biological information [141]. Starting with

the GO-terms at the highest level, the prioritizing algorithm iteratively applies this logic, searching for the (shortest) path to the root node. Along these shortest paths, GO terms are omitted. In summary, GOxploreR is available as an R package, allowing it to be easily integrated into existing analysis pipelines involving GO-terms.

5.1.2 Interpretation issues with prognostic biomarkers (II and III)

The main objective of Publication II and III [8, 9] is to systematically scrutinize the biological meaning of prognostic signatures of breast and prostate cancer respectively. We searched the literature for prognostic biomarkers and found that such signature genes are used in two interdependent ways. We refer to these as the predictive utility and biological utility of biomarkers. A brief definition of the two terms are as follows:

1. Predictive utility: The predictive utility of prognostic biomarkers means that biomarkers are used to categorize patients according to their prognostic state.
2. Biological utility: The biological utility of prognostic biomarkers means that biomarkers are used to provide biological insights into disease etiology.

The predictive ability of BM signatures is not disputed in this dissertation because the biomarkers can give an accurate prediction of cancer outcome. The biological utility, however, is challenged. If one thinks about it, both of these applications seem to be perfectly natural; after all, how could biomarkers with predictive utility not be useful for a biological interpretation of a disorder? Statistically, however, it is understood that there are two types of models. One of which is referred to as an explanatory or causal model, while the other is called the predictive model [142, 143]. Certainly, an explanatory model is more insightful than a predictive model since, although both may make predictions, only an explanatory model gives a reasonable justification for the fundamental mechanism about which the predictions are produced. A causal Bayesian network is an example of an explanatory model. As the results of Publication II and III show [8, 9], the biomarkers studied in this dissertation are no causal models. Therefore, the last-mentioned category of models are often referred to as black-box models [144].

Previous research on prognostic biomarkers has shown that such a disparity is also indispensable for biomarkers. Particularly, the authors of [61] examined 48

prognostic signatures of breast cancer, originating from independent, dedicated studies, and demonstrated that randomly chosen genes can also have identical predictive capabilities as the original signatures. Similar findings were previously reported in Ein-Dor et al. [145] where genes were ranked according to their correlation with survival outcome, and patients were classified using successive (non-overlapping) classes of genes. Even though the likelihood of discovering certain groups of correlated genes for more distant groups is decreased, the authors showed that such groups exist even for genes that do not rank at the top. Nonetheless, regardless of the selection mechanism of the various studies, all of them showed that there exist sets of genes that behave equally in the predictive task.

The work in Publication II and III [8, 9] has extended the above findings by acknowledging that disorders such as cancer represent complex rather than Mendelian diseases, the validity of the biological significance of biomarkers have been further tested in this dissertation through a novel GRP that entails removing signatures genes as well as genes involved in the same biological process as these genes. Prior studies on prognostic signatures have failed to do this, posing a major limitation to their studies because it is possible to unknowingly select the signature genes as random gene set. Furthermore, no removal of biological meaning is achieved by their study whereas the procedure in Publication II and III performed a strong removal of biological meaning, which enabled the conclusion made in this dissertation to be reached. This gene removal procedure can lead to a further reduction of the size of the pool of selectable genes, yet the results in Publication II and III showed that even among these remaining genes there exist random gene sets that perform equivalently in the prognostic predictive task. Importantly, owing to the absence of all genes that share biological processes with the original signature genes, the remaining genes do not share any biological meaning with such a signature. As this is true for every random gene set derived from the remaining genes, it is clear that these random gene sets have a biological meaning distinct from the original signature genes.

In summary, the results in publication II and III [8, 9] have shown that there is no justification for the dual utilization of prognostic biomarkers, i.e. for predictive and biological utility. In particular, the analysis systematically removed the risk that random gene sets would inadvertently have the same biological meaning as the original BM signatures genes by removing all those genes from the available pool of selectable genes. Returning to the statistical differentiation of the models mentioned

above, we conclude that none of the methods used to identify prognostic biomarkers in the studies investigated have established biological utility. Instead, they all form predictive utility that do not allow conclusions to be drawn about the underlying biology.

5.2 Conclusion

Prognostic biomarkers have been extensively investigated for decades due to their clinical usefulness in assisting patients. However, their general interpretation remained so far unclear. In this thesis, we focused on studying the interpretability of prognostics biomarkers of breast and prostate cancer. In order to be able to perform our analyses, we created a tool called GOxploreR (Publication I) [7] for the graph-based exploitation of GO. Due to the absence of software tools to directly explore the GO-DAG from a graphical theoretical viewpoint, our R package is beneficial since it complements currently available non-structural analysis tools.

Our results from Publication II and III [8, 9] raise awareness of a new issue relating to the biological meaning of prognostic biomarkers. To avoid misunderstanding and misuse of biomarkers, studies should focus on the predictive utility of the biomarkers and desist from establishing causal associations to disorders. The prognostic biomarkers that we have analyzed in this dissertation have no biological significance and are purely of predictive utility. Interestingly, it should be noted that even though these BM signatures lack a sensible biological meaning they are still useful in a clinical settings.

In order to help direct future studies on prognostic signatures, we recommend that any study claiming to have discovered prognostic signatures with a biological utility should use our GRP proposed in Publication II [8]. This provides a more stringent criterion than currently used in the literature to safeguard against false claims. Furthermore, this will aid in distinguishing between the predictive utility and the biological utility of prognostic biomarkers. Due to the fact that biomarkers are typically used in clinical settings any reduction in confusing aspects should be welcome for obtaining a clear understanding of biomarkers. Finally, prognostic biomarkers should be presented in a way that emphasizes their demonstrated ability to make predictions about the prognostic of patients.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs and A. Jemal. Cancer Statistics, 2021. *CA: a cancer journal for clinicians* 71.1 (2021), 7–33.

[2] M. Abdel-Rahman, D. Stockton, B. Rachet, T. Hakulinen and M. Coleman. What if cancer survival in Britain were the same as in Europe: how many deaths are avoidable?: *British journal of cancer* 101.2 (2009), S115–S124.

[3] K. Shyamala, H. Girish and S. Murgod. Risk of tumor cell seeding through biopsy and aspiration cytology. *Journal of International Society of Preventive & Community Dentistry* 4.1 (2014), 5.

[4] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research* 3.Mar (2003), 1157–1182.

[5] D. Ding, S. Han, H. Zhang, Y. He and Y. Li. Predictive biomarkers of colorectal cancer. *Computational biology and chemistry* 83 (2019), 107106.

[6] L. Ein-Dor, O. Zuk and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences* 103.15 (2006), 5923–5928.

[7] K. Manjang, S. Tripathi, O. Yli-Harja, M. Dehmer and F. Emmert-Streib. Graph-based exploitation of gene ontology using GOxploreR for scrutinizing biological significance. *Scientific reports* 10.1 (2020), 1–16. DOI: 10.1038/s41598-020-73326-3.

[8] K. Manjang, S. Tripathi, O. Yli-Harja, M. Dehmer, G. Glazko and F. Emmert-Streib. Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. *Scientific Reports* 11.1 (2021), 1–18. DOI: 10.1038/s41598-020-79375-y.

[9] K. Manjang, O. Yli-Harja, M. Dehmer and F. Emmert-Streib. Limitations of explainability for established prognostic biomarkers of prostate cancer. *Frontiers in Genetics* 12 (2021).

[10] L. Terese Winslow. *Medical And Scientific Illustration*. `https://www.teresewinslow.com/`. Accessed: 2021-01-26.

[11] *GLOBOCAN 2020*. `https://gco.iarc.fr/`. Accessed: 2021-01-28.

[12] D. R. Youlden, S. M. Cramb, N. A. Dunn, J. M. Muller, C. M. Pyke and P. D. Baade. The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality. *Cancer epidemiology* 36.3 (2012), 237–248.

[13] R. L. Siegel, K. D. Miller and A. Jemal. Cancer Statistics, 2018. *CA: a cancer journal for clinicians* (2018).

[14] F. Alkabban and T. Ferguson. Breast Cancer, StatPearls, StatPearls Publishing Copyright© 2020. (2020).

[15] A. Jemal, M. M. Center, C. DeSantis and E. M. Ward. Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiology and Prevention Biomarkers* 19.8 (2010), 1893–1907.

[16] S. Masood. Breast cancer subtypes: morphologic and biologic characterization. *Women's Health* 12.1 (2016), 103–119.

[17] T. Saphner, D. C. Tormey and R. Gray. Annual hazard rates of recurrence for breast cancer after primary therapy. *Journal of clinical oncology* 14.10 (1996), 2738–2746.

[18] K. J. Ready and B. K. Arun. Genetic predisposition to breast cancer and genetic counseling and testing. *Breast Cancer 2nd edition*. Springer, 2008, 57–81.

[19] Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, W. Ding et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 266.5182 (1994), 66–71.

[20] R. Wooster, G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, G. Micklem et al. Identification of the breast cancer susceptibility gene BRCA2. *Nature* 378.6559 (1995), 789–792.

[21] Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao and H.-P. Zhu. Risk factors and preventions of breast cancer. *International journal of biological sciences* 13.11 (2017), 1387.

[22] O. Yersal and S. Barutca. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology* 5.3 (2014), 412.

[23] A. H. Sims, A. Howell, S. J. Howell and R. B. Clarke. Origins of breast cancer subtypes and therapeutic implications. *Nature Clinical Practice Oncology* 4.9 (2007), 516–525.

[24] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* 98.19 (2001), 10869–10874.

[25] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen et al. Molecular portraits of human breast tumours. *nature* 406.6797 (2000), 747–752.

[26] K. D. Voduc, M. C. Cheang, S. Tyldesley, K. Gelmon, T. O. Nielsen and H. Kennecke. Breast cancer subtypes and the risk of local and regional relapse. *Journal of clinical oncology* 28.10 (2010), 1684–1691.

[27] R. Haque, S. A. Ahmed, G. Inzhakova, J. Shi, C. Avila, J. Polikoff, L. Bernstein, S. M. Enger and M. F. Press. Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiology and Prevention Biomarkers* 21.10 (2012), 1848–1855.

[28] A. W. Kurian, K. Fish, S. J. Shema and C. A. Clarke. Lifetime risks of specific breast cancer subtypes among women in four racial/ethnic groups. *Breast Cancer Research* 12.6 (2010), 1–9.

[29] C. A. Clarke, T. H. Keegan, J. Yang, D. J. Press, A. W. Kurian, A. H. Patel and J. V. Lacey Jr. Age-specific incidence of breast cancer subtypes: understanding the black–white crossover. *Journal of the National Cancer Institute* 104.14 (2012), 1094–1101.

[30] C. K. Anders, L. A. Carey and H. J. Burstein. ER/PR negative, HER2-negative (triple-negative) breast cancer. *Waltham, MA, UpToDate* (2020).

[31] A. Godoy-Ortiz, A. Sanchez-Muñoz, M. R. Chica Parrado, M. Álvarez, N. Ribelles, A. Rueda Dominguez and E. Alba. Deciphering HER2 breast cancer disease: biological and clinical implications. *Frontiers in oncology* 9 (2019), 1124.

[32] H. R. Umbas, A. Afriansyah, A. R. A. H. Hamid and C. A. Mochtar. Prostate specific antigen (PSA) kinetic as a prognostic factor in metastatic prostate cancer receiving androgen deprivation therapy: Systematic review and meta-analysis [version 1; referees: 1 approved, 2 approved with reservations]. *F1000Research* 7 (2018), 246.

[33] J. K. des Bordes, D. S. Lopez, M. D. Swartz and R. J. Volk. Sociodemographic disparities in cure-intended treatment in localized prostate cancer. *Journal of racial and ethnic health disparities* 5.1 (2018), 104–110.

[34] R. Siegel, K. Miller and A. Jemal. Cancer statistics, 2020. *CA Cancer Journal for Clinicians* 70.1 (2020), 7–30.

[35] M. Quinn and P. Babb. Patterns and trends in prostate cancer incidence, survival, prevalence and mortality. Part I: international comparisons. *BJU international* 90.2 (2002), 162–173.

[36] A. R. Patel and E. A. Klein. Risk factors for prostate cancer. *Nature Clinical Practice Urology* 6.2 (2009), 87–95.

[37] *Understanding Prostate Changes: A Health Guide for Men*. `https://www.cancer.gov/types/prostate/understanding-prostate-changes`. Accessed: 2021-02-4.

[38] P. Kheirandish and F. Chinegwundoh. Ethnic differences in prostate cancer. *British journal of cancer* 105.4 (2011), 481–485.

[39] R. Eeles, D. Dearnaley, A. Ardern-Jones, R. Shearer, D. Easton, D. Ford, S. Edwards, A. Dowe and 1. collaborators. Familial prostate cancer: the evidence and the cancer research campaign/British prostate group (CRC/BPG) UK familial prostate cancer study. *British journal of urology* 79.S1 (1997), 8–14.

[40] E. Giovannucci, E. B. Rimm, G. A. Colditz, M. J. Stampfer, A. Ascherio, C. C. Chute and W. C. Willett. A prospective study of dietary fat and risk of prostate cancer. *JNCI: Journal of the National Cancer Institute* 85.19 (1993), 1571–1579.

[41] W. J. Aronson, R. J. Barnard, S. J. Freedland, S. Henning, D. Elashoff, P. M. Jardack, P. Cohen, D. Heber and N. Kobayashi. Growth inhibitory effect of low fat diet on prostate cancer cells: results of a prospective, randomized dietary intervention trial in men with prostate cancer. *The Journal of urology* 183.1 (2010), 345–350.

[42] V. Venkateswaran and L. H. Klotz. Diet and prostate cancer: mechanisms of action and implications for chemoprevention. *Nature Reviews Urology* 7.8 (2010), 442.

[43] *Signs and Symptoms of Prostate Cancer*. `https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/signs-symptoms.html`. Accessed: 2021-02-5.

[44] S. J. Ruberg and L. Shen. Personalized medicine: four perspectives of tailored medicine. *Statistics in Biopharmaceutical Research* 7.3 (2015), 214–229.

[45] K. Sechidis, K. Papangelou, P. D. Metcalfe, D. Svensson, J. Weatherall and G. Brown. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics* 34.19 (2018), 3365–3376.

[46] E. A. Rakha, M. E. El-Sayed, A. R. Green, A. H. Lee, J. F. Robertson and I. O. Ellis. Prognostic markers in triple-negative breast cancer. *Cancer* 109.1 (2007), 25–32.

[47] F. Azuaje, Y. Devaux and D. Wagner. Computational biology for cardiovascular biomarker discovery. *Briefings in bioinformatics* 10.4 (2009), 367–377.

[48] T. Terkelsen, A. Krogh and E. Papaleo. CAncer bioMarker Prediction Pipeline (CAMPP)—A standardized framework for the analysis of quantitative biological data. *PLoS computational biology* 16.3 (2020), e1007665.

[49] E. A. Pudova, E. N. Lukyanova, K. M. Nyushko, D. S. Mikhaylenko, A. R. Zaretsky, A. V. Snezhkina, M. V. Savvateeva, A. A. Kobelyatskaya, N. V. Melnikova, N. N. Volchenko et al. Differentially expressed genes associated with prognosis in locally advanced lymph node-negative prostate cancer. *Frontiers in genetics* 10 (2019), 730.

[50] V. Kulasingam and E. P. Diamandis. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature clinical practice Oncology* 5.10 (2008), 588–599.

[51] W. Xu, Y. Wang, Y. Wang, S. Lv, X. Xu and X. Dong. Screening of differentially expressed genes and identification of NUF2 as a prognostic marker in breast cancer. *International journal of molecular medicine* 44.2 (2019), 390–404.

[52] Y. Cheng, K. Wang, L. Geng, J. Sun, W. Xu, D. Liu, S. Gong and Y. Zhu. Identification of candidate diagnostic and prognostic biomarkers for pancreatic carcinoma. *EBioMedicine* 40 (2019), 382–393.

[53] V. S. A. Jayanthi, A. B. Das and U. Saxena. Grade-specific diagnostic and prognostic biomarkers in breast cancer. *Genomics* 112.1 (2020), 388–396.

[54] Q. Zhang, X. Yin, Z. Pan, Y. Cao, S. Han, G. Gao, Z. Gao, Z. Pan and W. Feng. Identification of potential diagnostic and prognostic biomarkers for prostate cancer. *Oncology letters* 18.4 (2019), 4237–4245.

[55] K. Jiang, H. Liu, D. Xie and Q. Xiao. Differentially expressed genes ASPN, COL1A1, FN1, VCAN and MUC5AC are potential prognostic biomarkers for gastric cancer. *Oncology letters* 17.3 (2019), 3191–3202.

[56] M. Mourad, S. Moubayed, A. Dezube, Y. Mourad, K. Park, A. Torreblanca-Zanca, J. S. Torrecilla, J. C. Cancilla and J. Wang. Machine learning and feature selection applied to SEER data to reliably assess thyroid cancer prognosis. *Scientific reports* 10.1 (2020), 1–11.

[57] B. Shin, S. Park, J. H. Hong, H. J. An, S. H. Chun, K. Kang, Y.-H. Ahn, Y. H. Ko and K. Kang. Cascaded Wx: A novel prognosis-related feature selection framework in human lung adenocarcinoma transcriptomes. *Frontiers in genetics* 10 (2019), 662.

[58] J. Zhou, L. Li, L. Wang, X. Li, H. Xing and L. Cheng. Establishment of a SVM classifier to predict recurrence of ovarian cancer. *Molecular medicine reports* 18.4 (2018), 3589–3598.

[59] J. Zhi, J. Sun, Z. Wang and W. Ding. Support vector machine classifier for prediction of the metastasis of colorectal cancer. *International journal of molecular medicine* 41.3 (2018), 1419–1426.

[60] Y. Jiang, J. Xie, Z. Han, W. Liu, S. Xi, L. Huang, W. Huang, T. Lin, L. Zhao, Y. Hu et al. Immunomarker support vector machine classifier for prediction of gastric cancer survival and adjuvant chemotherapeutic benefit. *Clinical Cancer Research* 24.22 (2018), 5574–5584.

[61] D. Venet, J. E. Dumont and V. Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 7.10 (2011), e1002240.

[62] M. S. Schröder, A. C. Culhane, J. Quackenbush and B. Haibe-Kains. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* 27.22 (2011), 3206–3208.

[63] C. Brueffer, J. Vallon-Christersson, D. Grabau, A. Ehinger, J. Häkkinen, C. Hegardt, J. Malina, Y. Chen, P.-O. Bendahl, J. Manjer et al. Clinical value of RNA sequencing–based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter sweden cancerome analysis network—breast initiative. *JCO Precision Oncology* 2 (2018), 1–18.

[64] J. Liu, T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, A. J. Kovatich, C. C. Benz, D. A. Levine, A. V. Lee et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173.2 (2018), 400–416.

[65] M. C. Abba, E. Lacunza, M. Butti and C. M. Aldaz. Breast cancer biomarker discovery in the functional genomic age: a systematic review of 42 gene expression signatures. *Biomarker insights* 5 (2010), BMI–S5740.

[66] M. Adorno, M. Cordenonsi, M. Montagner, S. Dupont, C. Wong, B. Hann, A. Solari, S. Bobisse, M. B. Rondina, V. Guzzardo et al. A mutant-p53/Smad complex opposes p63 to empower TGFβ-induced metastasis. *Cell* 137.1 (2009), 87–98.

[67] I. Ben-Porath, M. W. Thomson, V. J. Carey, R. Ge, G. W. Bell, A. Regev and R. A. Weinberg. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nature genetics* 40.5 (2008), 499.

[68] M. Buess, D. S. Nuyten, T. Hastie, T. Nielsen, R. Pesich and P. O. Brown. Characterization of heterotypic interaction effects in vitro to deconvolute global gene expression profiles in cancer. *Genome biology* 8.9 (2007), 1–17.

[69] F. Buffa, A. Harris, C. West and C. Miller. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *British journal of cancer* 102.2 (2010), 428–435.

[70] S. L. Carter, A. C. Eklund, I. S. Kohane, L. N. Harris and Z. Szallasi. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature genetics* 38.9 (2006), 1043–1048.

[71] H. Y. Chang, J. B. Sneddon, A. A. Alizadeh, R. Sood, R. B. West, K. Montgomery, J.-T. Chi, M. Van De Rijn, D. Botstein and P. O. Brown. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* 2.2 (2004), e7.

[72] J.-T. Chi, Z. Wang, D. S. Nuyten, E. H. Rodriguez, M. E. Schaner, A. Salim, Y. Wang, G. B. Kristensen, Å. Helland, A.-L. Børresen-Dale et al. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS Med* 3.3 (2006), e47.

[73] N. P. Crawford, J. Alsarraj, L. Lukes, R. C. Walker, J. S. Officewala, H. H. Yang, M. P. Lee, K. Ozato and K. W. Hunter. Bromodomain 4 activation predicts breast cancer survival. *Proceedings of the National Academy of Sciences* 105.17 (2008), 6380–6385.

[74] H. Dai, L. van't Veer, J. Lamb, Y. D. He, M. Mao, B. M. Fine, R. Bernards, M. van de Vijver, P. Deutsch, A. Sachs et al. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer research* 65.10 (2005), 4059–4066.

[75] G. V. Glinsky, O. Berezovska, A. B. Glinskii et al. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *The Journal of clinical investigation* 115.6 (2005), 1503–1521.

[76] T. C. Hallstrom, S. Mori and J. R. Nevins. An E2F1-dependent gene expression program that determines the balance between proliferation and cell death. *Cancer cell* 13.1 (2008), 11–22.

[77] M. He, D. P. Mangiameli, S. Kachala, K. Hunter, J. Gillespie, X. Bian, H.-C. J. Shen and S. K. Libutti. Expression signature developed from a complex series of mouse models accurately predicts human breast cancer survival. *Clinical Cancer Research* 16.1 (2010), 249–259.

[78] G. Hu, R. A. Chong, Q. Yang, Y. Wei, M. A. Blanco, F. Li, M. Reiss, J. L.-S. Au, B. G. Haffty and Y. Kang. MTDH activation by 8q22 genomic gain promotes chemoresistance and metastasis of poor-prognosis breast cancer. *Cancer cell* 15.1 (2009), 9–20.

[79] S. Hua, R. Kittler and K. P. White. Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell* 137.7 (2009), 1259–1271.

[80] A. V. Ivshina, J. George, O. Senko, B. Mow, T. C. Putti, J. Smeds, T. Lindahl, Y. Pawitan, P. Hall, H. Nordgren et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research* 66.21 (2006), 10292–10301.

[81] M. Kok, R. H. Koornstra, T. C. Margarido, R. Fles, N. J. Armstrong, S. C. Linn, L. J. Van 't Veer and B. Weigelt. Mammosphere-derived gene set predicts outcome in patients with ER-positive breast cancer. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 218.3 (2009), 316–326.

[82] J. E. Korkola, E. Blaveri, S. DeVries, D. H. Moore, E. S. Hwang, Y.-Y. Chen, A. L. Estep, K. L. Chew, R. H. Jensen and F. M. Waldman. Identification of a robust gene signature that predicts breast cancer outcome in independent data sets. *BMC cancer* 7.1 (2007), 1–13.

[83] R. Liu, X. Wang, G. Y. Chen, P. Dalerba, A. Gurney, T. Hoey, G. Sherlock, J. Lewicki, K. Shedden and M. F. Clarke. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *New England Journal of Medicine* 356.3 (2007), 217–226.

[84] X.-J. Ma, R. Salunga, J. T. Tuggle, J. Gaudet, E. Enright, P. McQuary, T. Payette, M. Pistone, K. Stecker, B. M. Zhang et al. Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences* 100.10 (2003), 5974–5979.

[85] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences* 102.38 (2005), 13550–13555.

[86] S. Mori, J. T. Chang, E. R. Andrechek, N. Matsumura, T. Baba, G. Yao, J. W. Kim, M. Gatza, S. Murphy and J. R. Nevins. Anchorage-independent cell growth signature identifies tumors with metastatic potential. *Oncogene* 28.31 (2009), 2796–2805.

[87] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine* 351.27 (2004), 2817–2826.

[88] S. Calza, P. Hall, G. Auer, J. Björhle, S. Klaar, U. Kronenwett, E. T. Liu, L. Miller, A. Ploner, J. Smeds et al. Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Research* 8.4 (2006), 1–9.

[89] X.-H. Pei, F. Bai, M. D. Smith, J. Usary, C. Fan, S.-Y. Pai, I.-C. Ho, C. M. Perou and Y. Xiong. CDK inhibitor p18INK4c is a downstream target of GATA3 and restrains mammary luminal progenitor cell proliferation and tumorigenesis. *Cancer cell* 15.5 (2009), 389–401.

[90] S. Ramaswamy, K. N. Ross, E. S. Lander and T. R. Golub. A molecular signature of metastasis in primary solid tumors. *Nature genetics* 33.1 (2003), 49–54.

[91] J. A. Reuter, S. Ortiz-Urda, M. Kretz, J. Garcia, F. A. Scholl, A. M. Pasmooij, D. Cassarino, H. Y. Chang and P. A. Khavari. Modeling inducible human tissue neoplasia identifies an extracellular matrix interaction network involved in cancer progression. *Cancer cell* 15.6 (2009), 477–488.

[92] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey and A. M. Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences* 101.25 (2004), 9309–9314.

[93] L. H. Saal, P. Johansson, K. Holm, S. K. Gruvberger-Saal, Q.-B. She, M. Maurer, S. Koujak, A. A. Ferrando, P. Malmström, L. Memeo et al. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proceedings of the National Academy of Sciences* 104.18 (2007), 7564–7569.

[94] M. Shipitsin, L. L. Campbell, P. Argani, S. Weremowicz, N. Bloushtain-Qimron, J. Yao, T. Nikolskaya, T. Serebryiskaya, R. Beroukhim, M. Hu et al. Molecular definition of breast tumor heterogeneity. *Cancer cell* 11.3 (2007), 259–273.

[95] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the national academy of sciences* 100.14 (2003), 8418–8423.

[96] C. Sotiriou, S.-Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris and E. T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences* 100.18 (2003), 10393–10398.

[97] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute* 98.4 (2006), 262–272.

[98] X. Ge, S. Yamamoto, S. Tsutsumi, Y. Midorikawa, S. Ihara, S. M. Wang and H. Aburatani. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* 86.2 (2005), 127–141.

[99] J. H. Taube, J. I. Herschkowitz, K. Komurov, A. Y. Zhou, S. Gupta, J. Yang, K. Hartwell, T. T. Onder, P. B. Gupta, K. W. Evans et al. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proceedings of the National Academy of Sciences* 107.35 (2010), 15449–15454.

[100] S. F. Tavazoie, C. Alarcón, T. Oskarsson, D. Padua, Q. Wang, P. D. Bos, W. L. Gerald and J. Massagué. Endogenous human microRNAs that suppress breast cancer metastasis. *Nature* 451.7175 (2008), 147–152.

[101] S. Valastyan, F. Reinhardt, N. Benaich, D. Calogrias, A. M. Szász, Z. C. Wang, J. E. Brock, A. L. Richardson and R. A. Weinberg. *RETRACTED: a pleiotropically acting microRNA, miR-31, inhibits breast cancer metastasis*. 2009.

[102] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* 347.25 (2002), 1999–2009.

[103] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* 365.9460 (2005), 671–679.

[104] S. E. Wang, B. Xiang, M. Guix, M. G. Olivares, J. Parker, C. H. Chung, A. Pandiella and C. L. Arteaga. Transforming growth factor β engages TACE and ErbB3 to activate phosphatidylinositol-3 kinase/Akt in ErbB2-overexpressing breast cancer and desensitizes cells to trastuzumab. *Molecular and cellular biology* 28.18 (2008), 5605–5620.

[105] A. L. Welm, J. B. Sneddon, C. Taylor, D. S. Nuyten, M. J. van de Vijver, B. H. Hasegawa and J. M. Bishop. The macrophage-stimulating protein pathway promotes metastasis in a mouse model for breast cancer and predicts poor prognosis in humans. *Proceedings of the National Academy of Sciences* 104.18 (2007), 7570–7575.

[106] R. B. West, D. S. Nuyten, S. Subramanian, T. O. Nielsen, C. L. Corless, B. P. Rubin, K. Montgomery, S. Zhu, R. Patel, T. Hernandez-Boussard et al. Determination of stromal signatures in breast carcinoma. *PLoS Biol* 3.6 (2005), e187.

[107] M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* 13.6 (2002), 1977–2000.

[108] D. J. Wong, H. Liu, T. W. Ridky, D. Cassarino, E. Segal and H. Y. Chang. Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell stem cell* 2.4 (2008), 333–344.

[109] D. J. Wong, D. S. Nuyten, A. Regev, M. Lin, A. S. Adler, E. Segal, M. J. van de Vijver and H. Y. Chang. Revealing targeted therapy for human cancer by gene module maps. *Cancer research* 68.2 (2008), 369–378.

[110] J. Yu, J. Yu, D. R. Rhodes, S. A. Tomlins, X. Cao, G. Chen, R. Mehra, X. Wang, D. Ghosh, R. B. Shah et al. A polycomb repression signature in metastatic prostate cancer predicts cancer outcome. *Cancer research* 67.22 (2007), 10657–10663.

[111] L. Agell, S. Hernández, L. Nonell, M. Lorenzo, E. Puigdecanet, S. de Muga, N. Juanpere, R. Bermudo, P. L. Fernández, J. A. Lorente et al. A 12-gene expression signature is associated with aggressive histological in prostate cancer: SEC14L1 and TCEB1 genes are potential markers of progression. *The American journal of pathology* 181.5 (2012), 1585–1594.

[112] M. Bibikova, E. Chudin, A. Arsanjani, L. Zhou, E. W. Garcia, J. Modder, M. Kostelec, D. Barker, T. Downs, J.-B. Fan et al. Expression signatures that correlated with Gleason score and relapse in prostate cancer. *Genomics* 89.6 (2007), 666–672.

[113] T. A. Bismar, F. Demichelis, A. Riva, R. Kim, S. Varambally, L. He, J. Kutok, J. C. Aster, J. Tang, R. Kuefer et al. Defining aggressive prostate cancer using a 12-gene model. *Neoplasia (New York, NY)* 8.1 (2006), 59.

[114] X. Chen, J. Wang, X. Peng, K. Liu, C. Zhang, X. Zeng and Y. Lai. Comprehensive analysis of biomarkers for prostate cancer based on weighted gene co-expression network analysis. *Medicine* 99.14 (2020), e19628.

[115] X. Chen, S. Xu, M. McClelland, F. Rahmatpanah, A. Sawyers, Z. Jia and D. Mercola. An accurate prostate cancer prognosticator using a seven-gene signature plus Gleason score and taking cell type heterogeneity into account. *PloS one* 7.9 (2012), e45178.

[116] J. C. Cheville, R. J. Karnes, T. M. Therneau, F. Kosari, J.-M. Munz, L. Tillmans, E. Basal, L. J. Rangel, E. Bergstralh, I. V. Kovtun et al. Gene panel model predictive of outcome in men at high-risk of systemic progression and death from prostate cancer after radical retropubic prostatectomy. *Journal of Clinical Oncology* 26.24 (2008), 3930.

[117] J. Chu, N. Li and W. Gai. Identification of genes that predict the biochemical recurrence of prostate cancer. *Oncology letters* 16.3 (2018), 3447–3452.

[118] J. Cuzick, G. P. Swanson, G. Fisher, A. R. Brothman, D. M. Berney, J. E. Reid, D. Mesher, V. Speights, E. Stankiewicz, C. S. Foster et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *The lancet oncology* 12.3 (2011), 245–255.

[119] S. Irshad, M. Bansal, M. Castillo-Martin, T. Zheng, A. Aytes, S. Wenske, C. Le Magnen, P. Guarnieri, P. Sumazin, M. C. Benson et al. A molecular signature predictive of indolent prostate cancer. *Science translational medicine* 5.202 (2013), 202ra122–202ra122.

[120] S. Larkin, S. Holmes, I. Cree, T. Walker, V. Basketter, B. Bickers, S. Harris, S. D. Garbis, P. Townsend and C. Aukim-Hastie. Identification of markers of prostate cancer progression using candidate gene expression. *British journal of cancer* 106.1 (2012), 157–165.

[121] F. Li, J.-P. Ji, Y. Xu and R.-L. Liu. Identification a novel set of 6 differential expressed genes in prostate cancer that can potentially predict biochemical recurrence after curative surgery. *Clinical and Translational Oncology* 21.8 (2019), 1067–1075.

[122] Q. Long, B. A. Johnson, A. O. Osunkoya, Y.-H. Lai, W. Zhou, M. Abramovitz, M. Xia, M. B. Bouzyk, R. K. Nam, L. Sugar et al. Protein-coding and microRNA biomarkers of recurrence of prostate cancer following radical prostatectomy. *The American journal of pathology* 179.1 (2011), 46–54.

[123] T. Nakagawa, T. M. Kollmeyer, B. W. Morlan, S. K. Anderson, E. J. Bergstralh, B. J. Davis, Y. W. Asmann, G. G. Klee, K. V. Ballman and R. B. Jenkins. A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. *PloS one* 3.5 (2008), e2318.

[124] K. L. Penney, J. A. Sinnott, K. Fall, Y. Pawitan, Y. Hoshida, P. Kraft, J. R. Stark, M. Fiorentino, S. Perner, S. Finn et al. mRNA expression signature of Gleason grade predicts lethal prostate cancer. *Journal of Clinical Oncology* 29.17 (2011), 2391.

[125] G. K. Reddy and S. P. Balk. Clinical utility of microarray-derived genetic signatures in predicting outcomes in prostate cancer. *Clinical genitourinary cancer* 5.3 (2006), 187–189.

[126] H. Ross-Adams, A. Lamb, M. Dunning, S. Halim, J. Lindberg, C. Massie, L. Egevad, R. Russell, A. Ramos-Montoya, S. Vowler et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study. *EBioMedicine* 2.9 (2015), 1133–1144.

[127] R. W. Ross, M. D. Galsky, H. I. Scher, J. Magidson, K. Wassmann, G.-S. M. Lee, L. Katz, S. K. Subudhi, A. Anand, M. Fleisher et al. A whole-blood RNA transcript-based prognostic model in men with castration-resistant prostate cancer: a prospective study. *The lancet oncology* 13.11 (2012), 1105–1113.

[128] N. L. Sharma, C. E. Massie, A. Ramos-Montoya, V. Zecchini, H. E. Scott, A. D. Lamb, S. MacArthur, R. Stark, A. Y. Warren, I. G. Mills et al. The androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. *Cancer cell* 23.1 (2013), 35–47.

[129] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell* 1.2 (2002), 203–209.

[130] Z. Song, Y. Huang, Y. Zhao, H. Ruan, H. Yang, Q. Cao, D. Liu, X. Zhang and K. Chen. The identification of potential biomarkers and biological pathways in prostate cancer. *Journal of Cancer* 10.6 (2019), 1398.

[131] A. J. Stephenson, A. Smith, M. W. Kattan, J. Satagopan, V. E. Reuter, P. T. Scardino and W. L. Gerald. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer: Interdisciplinary International Journal of the American Cancer Society* 104.2 (2005), 290–298.

[132] D. Talantov, T. A. Jatkoe, M. Böhm, Y. Zhang, A. M. Ferguson, P. D. Stricker, M. W. Kattan, R. L. Sutherland, J. G. Kench, Y. Wang et al. Gene based prediction of clinically localized prostate cancer progression after radical prostatectomy. *The Journal of urology* 184.4 (2010), 1521–1528.

[133] D. G. Tandefelt, J. L. Boormans, H. A. van der Korput, G. W. Jenster and J. Trapman. A 36-gene signature predicts clinical progression in a subgroup of ERG-positive prostate cancers. *European urology* 64.6 (2013), 941–950.

[134] L. True, I. Coleman, S. Hawley, C.-Y. Huang, D. Gifford, R. Coleman, T. M. Beer, E. Gelmann, M. Datta, E. Mostaghel et al. A molecular correlate to the Gleason grading system for prostate adenocarcinoma. *Proceedings of the National Academy of Sciences* 103.29 (2006), 10991–10996.

[135] L.-Y. Wang, J.-J. Cui, T. Zhu, W.-H. Shao, Y. Zhao, S. Wang, Y.-P. Zhang, J.-C. Wu and L. Zhang. Biomarkers identified for prostate cancer patients through genome-scale screening. *Oncotarget* 8.54 (2017), 92055.

[136] C.-L. Wu, B. E. Schroeder, X.-J. Ma, C. J. Cutie, S. Wu, R. Salunga, Y. Zhang, M. W. Kattan, C. A. Schnabel, M. G. Erlander et al. Development and validation of a 32-gene prognostic index for prostate cancer progression. *Proceedings of the National Academy of Sciences* 110.15 (2013), 6121–6126.

[137] G. O. Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research* 47.D1 (2019), D330–D338.

[138] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig et al. Gene ontology: tool for the unification of biology. *Nature genetics* 25.1 (2000), 25–29.

[139] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: `https://www.R-project.org/`.

[140] F. Emmert-Streib and M. Dehmer. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction* 1.3 (2019), 1013–1038.

[141] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki. DAVID: database for annotation, visualization, and integrated discovery. *Genome biology* 4.9 (2003), 1–11.

[142] G. Shmueli et al. To explain or to predict?: *Statistical science* 25.3 (2010), 289–310.

[143] L. Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16.3 (2001), 199–231.

[144] F. Emmert-Streib, O. Yli-Harja and M. Dehmer. Explainable artificial intelligence and machine learning: A reality rooted perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.6 (2020), e1368.

[145] L. Ein-Dor, I. Kela, G. Getz, D. Givol and E. Domany. Outcome signature genes in breast cancer: is there a unique set?: *Bioinformatics* 21.2 (2005), 171–178.

PUBLICATIONS

PUBLICATION

I

Graph-based exploitation of gene ontology using GOxploreR for scrutinizing biological significance

K. Manjang, S. Tripathi, O. Yli-Harja, M. Dehmer and F. Emmert-Streib

Scientific reports 10.1 (2020), 1–16
DOI: 10.1038/s41598-020-73326-3

SCIENTIFIC REPORTS
nature research

OPEN

Graph-based exploitation of gene ontology using GOxploreR for scrutinizing biological significance

Kalifa Manjang[1], Shailesh Tripathi[1], Olli Yli-Harja[2,3,6], Matthias Dehmer[4,5] & Frank Emmert-Streib[1,6]✉

Gene ontology (GO) is an eminent knowledge base frequently used for providing biological interpretations for the analysis of genes or gene sets from biological, medical and clinical problems. Unfortunately, the interpretation of such results is challenging due to the large number of GO terms, their hierarchical and connected organization as directed acyclic graphs (DAGs) and the lack of tools allowing to exploit this structural information explicitly. For this reason, we developed the `R` package `GOxploreR`. The main features of `GOxploreR` are (I) easy and direct access to structural features of GO, (II) structure-based ranking of GO-terms, (III) mapping to reduced GO-DAGs including visualization capabilities and (IV) prioritizing of GO-terms. The underlying idea of `GOxploreR` is to exploit a graph-theoretical perspective of GO as manifested by its DAG-structure and the containing hierarchy levels for cumulating semantic information. That means all these features enhance the utilization of structural information of GO and complement existing analysis tools. Overall, `GOxploreR` provides exploratory as well as confirmatory tools for complementing any kind of analysis resulting in a list of GO-terms, e.g., from differentially expressed genes or gene sets, GWAS or biomarkers. Our `R` package `GOxploreR` is freely available from CRAN.

The gene ontology (GO) consortium funded by the National Institute of Health (NIH) started in 1998. Initially, GO contained only three model organisms but extended since then to over 3200[1,2]. The ontology is structured into three distinct aspects of gene function, namely, molecular function (MF), cellular component (CC), and biological process (BP) together with over 45, 000 terms and 130, 000 relations. However, the majority of information is centered around ten model organisms (human, mouse, rat, zebrafish, drosophila, *C. elegans*, *D. discoideum*, *S. cerevisiae*, *S. pombe*, *A. thalia* and *E. coli*)[2]. In addition, GO includes annotations by linking specific gene products to GO-terms. This allows the connection between genes and GO-terms for deriving organism-specific information. Currently, GO is the most comprehensive and widely used knowledge base concerning functional information about genes[3–6].

A reason for the widespread applicability of GO is its generality. That means instead of providing solutions to particular problems, GO provides generic information that can be connected to any list of genes or gene products regardless of the type of upstream analysis that generated such a list. For instance, investigations that can lead to a list of genes are from studies about differentially expressed genes or gene sets, GWAS (genome-wide association study), biomarkers or gene regulatory networks[7–13]. These studies could be of biological, medical, clinical or pharmacological nature making GO useful across the life and health sciences.

Interestingly, despite the widespread usage of GO for a number of different application types[7,14,15], for exploring the GO knowledge base from a graph theoretical perspective[16,17] the available tools are surprisingly sparse and

[1]Predictive Society and Data Analytics Lab, Tampere University, Tampere, Korkeakoulunkatu 10, 33720 Tampere, Finland. [2]Computational Systems Biology, Tampere University, Tampere, Korkeakoulunkatu 10, 33720 Tampere, Finland. [3]Institute for Systems Biology, Seattle, WA, USA. [4]Department of Biomedical Computer Science and Mechatronics, UMIT-The Health and Life Science University, 6060 Hall in Tyrol, Austria. [5]College of Artificial Intelligence, Nankai University, Tianjin 300350, China. [6]Institute of Biosciences and Medical Technology, Tampere University, Tampere, Korkeakoulunkatu 10, 33720 Tampere, Finland. ✉email: frank.emmert-streib@tuni.fi

only very basic functions are available for obtaining structural information[18–20]. However, no dedicated functions are ready-for-use that give us, e.g., information about the GO-level of a GO-term, the category (regular node, jump node or leaf node) of a GO-term, the adjacency matrix of the GO-DAG of BP terms or all GO-terms on a specific GO-level, to name just a few. Furthermore, existing tools do not provide means for reducing the overall complexity of GO that would be amenable, for instance, for a visualization. Given the size of GO containing thousands of GO-terms, such a simplification would be highly desirable.

For these reasons, we created the `R` package `GOxploreR` to fill this gap. Our package provides direct access to structural information allowing the efficient exploitation of graph-theoretical properties of a DAG (directed acyclic graph) for further analysis. We provide also information on a low level. For instance, given a list of Entrez Gene IDs our package includes an (online) function to provide the BP, MF or CC of GO-terms associated with these genes. To retrieve the most current GO-terms, we use the biomartR package to query the Ensembl website. However, for obtaining fast information, we added also an offline version of these functions with pre-assembled information. This functionality is supported for ten organisms.

Aside from functions for the quantification of structural properties of GO-DAGs, we provide also visualization capabilities. Due to the size of GO our visualizations aim at a simplified representation. Specifically, by categorizing GO-terms into three classes—called regular nodes (RN), jump nodes (JN) and leaf nodes (LN)—we obtain a simplified representation of a GO-DAG with at most three nodes on each GO-level and the connections among them. These categories simplify the semantic attributes of GO-terms significantly yet provide important information regarding their connectivity. In this way, the GO-DAG of human for BP with 29, 699 GO-terms is reduced to a simplified DAG with 39 nodes, which is amenable for a visualization. We provide also extensions of such a visualization by, e.g., filtering for a set of GO-terms. This leads to a further reductions of complexity and can be utilized for compact visualizations of large lists of significant genes, gene sets or pathways. Finally, we provide a function for prioritizing a list of GO-terms as obtained, e.g., from differentially expressed genes, that reflects the structural positions of these GO-terms and their biological-semantic importance within the entire GO-DAG.

In general, one of the main applications of GO is the identification of over- or under-represented GO-terms for a specified gene list (as a result, e.g., from identifying differentially expressed genes) utilizing a hypergeometric test (also known as Fisher's exact test)[21,22]. A problem with this is that GO has a hierarchical structure in the form of a directed acyclic graph (DAG), which means that the GO-terms are dependent on each other. However, the above approaches ignore this dependency structure. For compensating this omission, semantic measures have been suggested, e.g., utilizing frequencies to assess the similarity/distance between GO-terms[23]. Alternatively, information about the connection of GO-terms has been included to a certain degree for enrichment analysis, e.g.,[24]. Although such approaches are more informative, in practice, they are often ignored and the structure-less methods are preferred because they are simpler to apply and interpret. Another problem is that different semantic measures seem to be preferable for particular biological data and applications, which further complicates the selection of such measures enormously[25].

In contrast, the R package `GOxploreR` is different to the above approaches in the following way. Specifically its main features include (I) a direct access to structural features of GO, (II) a structure-based ranking of GO-terms, (III) a mapping from a GO-DAG to a reduced GO-DAG, (IV) a visualization of reducuded GO-DAGs and (V) an algorithm for prioritizing GO-terms. That means the providesd features are meant to complement, e.g., approaches for identifying enriched GO-terms by providing alternative approaches for the analysis of GO-terms. Overall, `GOxploreR` can help in improving some of the above discussed shortcomings by providing novel ways for graph-based exploitations of the GO knowledge base to simplify the interpretation of large sets of significant GO-terms by utilizing structural information from the underlying DAG. Due to the fact that such a list of GO-terms can come from any type of upstream analysis, `GOxploreR` is a very versatile and flexible tool with respect to potential applications in the life and health sciences.

This paper is organized as follows. In the next section, we describe the underlying methodology of `GOxploreR` and the provided functionality. Then we showcase the applicability of `GOxploreR` by highlighting some of its features and implemented functions. This paper finishes with a discussion of the available functions, a comparison to existing tools and concluding remarks.

Methods

In this section, we provide technical information about the main features provided by `GOxploreR`. First, we discuss how one obtains a directed acyclic graph (DAG) for given GO-terms. Then we discuss organism-specific GO-DAGs and a mapping that converts such a DAG into a reduced GO-DAG. Finally, we discuss an algorithm for prioritizing GO-terms.

Determining the GO-DAG. The problem with existing packages is that none provides a function to directly obtain a GO-DAG for a domain, i.e., BP, MF or CC, in the form of an adjacency matrix. Instead, they provide local information which needs to be used for *deducing* such a tree tediously. For instance, GOdb provides the function `GOBPCHILDREN` to get the children of a GO term for BP. For the other two domains similar functions are available. The problem is that a children node does not need to be on the next hierarchy level but can jump further down the DAG. For an example see Fig. 1. In this figure, the child of node 2 is node 8 which is located on level 4, i.e., the child jumps from level 1, the location of its parent, to level 4.

The following example demonstrates how one can deduce a GO-DAG from this information. First, we list all children of a GO term (as obtained via the command `GOBPCHILDREN`).

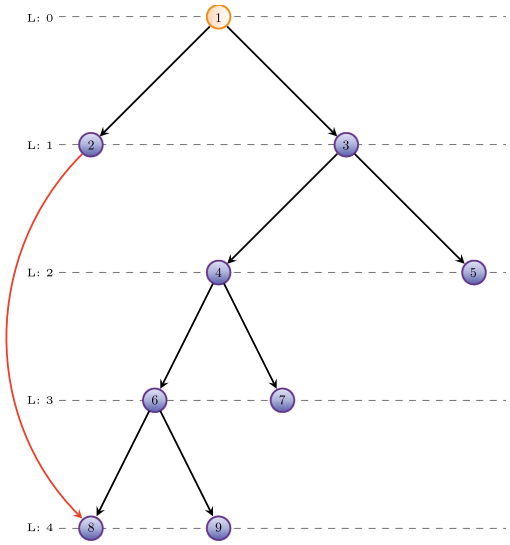


Figure 1. An example for a toy GO-DAG containing 9 GO-terms, whereas each node corresponds to one GO-term. The children of a node can jump over levels, as shown in red for the connection between node 2 and 8.

$$CH(x_1) = \{x_2, x_3\} \tag{1}$$

$$CH(x_2) = \{x_8\} \tag{2}$$

$$CH(x_3) = \{x_4, x_5\} \tag{3}$$

$$CH(x_4) = \{x_6, x_7\} \tag{4}$$

$$CH(x_5) = \emptyset \tag{5}$$

$$CH(x_6) = \{x_8, x_9\} \tag{6}$$

$$CH(x_7) = \emptyset \tag{7}$$

$$CH(x_8) = \emptyset \tag{8}$$

$$CH(x_9) = \emptyset \tag{9}$$

The root node is unique and we assign it the level 0, i.e., $L(x_1) = 0$. The children for the root node receive as first assignment for a level the value $L(x_1) + 1 = 1$, i.e.,

$$L(x_2) = \{1\} \tag{10}$$

$$L(x_3) = \{1\} \tag{11}$$

We wrote the right-hand side as a set because if such a node appeares again, we just add the new level value to this set. Going through the list of children, we assign each children of a node x_i the value $L(x_i) + 1$.

$$CH(x_2) \rightarrow L(x_8) = \{2\} \tag{12}$$

$$CH(x_3) \rightarrow L(x_4) = \{2\},\ L(x_5) = \{2\} \tag{13}$$

$$CH(x_4) \rightarrow L(x_6) = \{3\},\ L(x_7) = \{3\} \tag{14}$$

$$CH(x_6) \rightarrow L(x_8) = \{2, 4\},\ L(x_9) = \{4\} \tag{15}$$

From the last line we see that x_8 appears once on level 2 and once on level 4, which is correct if one looks at Fig. 1. However, there is just one correct level for x_8 and this is level 4. In general, if more than one level is assigned to a node then the correct one is the largest of these values.

Such a GO-DAG can be constructed for every domain, i.e., biological process, molecular function and cellular component. In our package, we call the resulting graphs:

- `g.GO-DAG.BP`: A DAG for all GO-terms of biological processes.
- `g.GO-DAG.MF`: A DAG for all GO-terms of molecular functions.
- `g.GO-DAG.CC`: A DAG for all GO-terms of cellular components.

Organism-specific GO-DAG. Starting from a GO-DAG for a domain, as constructed in the previous section and using a list of all genes from an organisms, we can map these genes to GO-terms. For a particular organism, not all GO-terms may be present but only a subset. Such a subset can then be mapped back to the entire GO-DAG of the knowledge base. This gives a subtree of the general GO-DAG that is organism-specific. Using the function `GetDAG(organism = o.name, domain = "BP")` one obtains, e.g., a GO-DAG of BPs for the organism given by 'o.name'. For all domains, the following functions can be used:

- `GetDAG(organism = o.name, domain = "BP")`: A sub-DAG for all GO-terms of biological processes for organism 'o.name'.
- `GetDAG(organism = o.name, domain = "MF")`: A sub-DAG for all GO-terms of molecular functions for organism 'o.name'.
- `GetDAG(organism = o.name, domain = "CC")`: A sub-DAG for all GO-terms of cellular components for organism 'o.name'.

Reduced GO-DAG. Visualizing one of the GO-DAGs determined above (for all GO-terms or for organism-specific GO-terms) is usually challenging because of the size of such graphs containing thousands of GO-terms corresponding to nodes in a graph. For this reason, we derive a simplified GO-DAG, containing only dozens of nodes, that can be easily visualized to obtain a global overview of all used GO-terms.

In order to simplify a GO-DAG, we introduce the following categorization of GO-terms, excluding the root node. This categorization is applied to each level separately:

- A GO-term is in category 'leaf node' (LN) if it has no children.
- A GO-term is in category 'regular node' (RN) if all its children are on the next level.
- A GO-term is in category 'jump node' (JN) if it has children and at least one of these is not on the next level.

We apply this categorization for all GO-terms. This results in the mapping

GO-term $X \rightarrow$ GO-term category on level L

That means we have functions of the form

$$(c, l) = f(X) \tag{16}$$

with $c \in \{\text{LN, RN, JN}\}$ and $l \in \mathbb{N}$. For instance, from Fig. 1 follows $3 \rightarrow$ RN on level 1 and $2 \rightarrow$ JN on level 1, which can be written formally as

$$(\text{ RN }, 1) = f(3) \tag{17}$$

$$(\text{ JN }, 1) = f(2) \tag{18}$$

Algorithmically, the implementation is described in 1.

Algorithm 1: CATEGORIZATION OF GO-TERMS

For a GO-DAG with L levels, M nodes, adjacency matrix $A \in \mathbb{R}^{M\times M}$ and level function $l = g(i)$ for $i \in \{1,\ldots,M\}$ and $l \in \{0,\ldots,L\}$
Initialize hash H # for nodes in GO-DAG
Initialize hash V # for nodes in simplified GO-DAG
Initialize hash F
Initialize matrix $C \in \mathbb{R}^{(L+1)\times 3}$
Initialize vectors Ca, ca, h
for $i \in \{1,\ldots,M\}$ do
    $S = links(A(i,))$ # find all nodes S linking from i (outgoing links from i)
    $l_i = g(i)$
    $K = /\ 0$
    foreach *node* $j \in S$ do
        $l_j = g(j)$ # find the level of node j
        $K \leftarrow l_j$
    if $S = /\ 0$ then
        $c_i =$LN
    else if l_j *exists in* K *with* $l_j > l_i + 1$ then
        $c_i =$JN
    else if $|S| > 0$ then
        $c_i =$RN
    set $H\{(c_i, l_i)\} \leftarrow i$ # store set of nodes i with c_i and l_i
    set $Ca(i) = c_i$ # categorize node i
$k = 1$ # node ID for nodes in simplified GO-DAG
for $l \in \{0,\ldots,L\}$ do
    # summarize nodes of the same category
    $C(l,1) = |H\{(\text{LN}, l)\}|$ # number of leaf nodes on level l
    $C(l,2) = |H\{(\text{RN}, l)\}|$
    $C(l,3) = |H\{(\text{JN}, l)\}|$
    foreach $C(l,c) > 0$ do
        set $V\{k\} = H\{(c,l)\}$ # mapping between old and new node IDs
        set $F\{(c,l)\} = k$
        set $h(k) = l$ # level function of simplified GO-DAG
        set $ca(k) = c$
        $k = k + 1$
$N = |V|$ # number of nodes in simplified GO-DAG

In addition to the node categorization, we need to find the connections between these nodes. This is realized via the implementation shown in Algorithm 2.

Algorithm 2: CALCULATE NUMBER OF LINKS BETWEEN CATEGORY NODES.

For A, F, M, N, h and Ca; see Algo 1
Initialize adjacency matrix B with $B \in \mathbb{R}^{N\times N}$ for simplified GO-DAG
for $l_1 \in \{0,\ldots,L\}$ do
    foreach *node* x *on level* l_1 do
        $c_1 = Ca(x)$ # find the category of node x
        $i_1 = F\{(c_1, l_1)\}$
        $S = Ch(x)$ # find all children of x using A
        foreach $y \in S$ do
            $c_2 = Ca(y)$ # find the category of node y
            $l_2 = g(y)$ # find the level of node y
            $i_2 = F\{(c_2, l_2)\}$
            $B(i_1, i_2) = B(i_1, i_2) + 1$

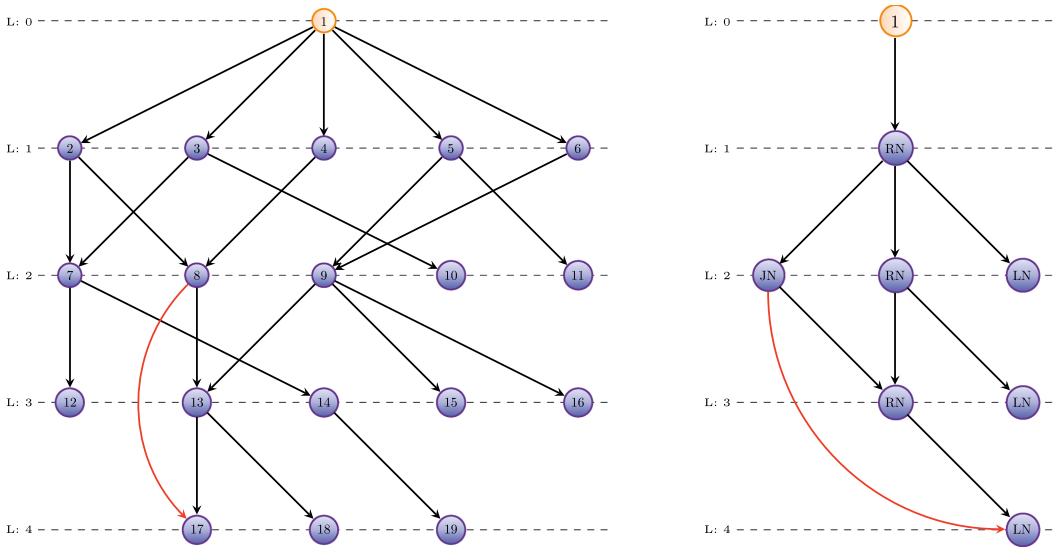


Figure 2. An example for the construction of a reduced GO-DAG. Left: An ordinary GO-DAG with 19 GO terms is shown. Right: The reduced GO-DAG with 8 nodes summarizes the left graph. Note, the nodes in the right graph are no GO-terms but node categories, i.e., either RN, JN or LN.

Overall, a GO-DAG is described by an adjacency matrix A and a level function g and analogously, a reduced GO-DAG is described by adjacency matrix B and level function h and C (number of original nodes summarized by a new category).

In Fig. 2 we show a complete example for this mapping. The GO-DAG on the left-hand side has 19 GO terms and the resulting simplified GO-DAG on the right-hand side has only 8 nodes, whereas these nodes correspond to the three GO categories (RN, JN & LN) defined above. As one can see, each level will contain at most 3 nodes because this is the number of different categories. However, it is possible to have even fewer nodes, if a category is absent on a level.

Importantly, this transformation can be applied to any GO-DAG, regardless if this DAG is for all GO terms of, e.g., BPs, or for an organism-specific GO-DAG.

Prioritizing lists of GO-terms. In general, the comparison of GO-terms with respect to their biological-semantic importance is complex. However, the comparison of GO-terms along a path is much simpler because the higher a level of a GO-term is the more specific is its biological information[26]. That means *vertically* one wants to traverse a DAG along a path as far down as possible. This implies that the GO-term at the end of a path is most interesting compared to all other GO-terms along this path. This increase in the semantic meaning along *vertical* paths is exploited by our algorithm for prioritizing lists of GO-terms.

Algorithm 3: PRIORITIZING A LIST OF GO-TERMS.

For a list, H, of GO-terms in domain XX, a GO-DAG of XX and level function g
Initialize a list R
$n = |H|$
foreach $i \in H$ do
    $l_i = g(i)$ # find level for each GO-term
while $n > 0$ do
    $r =$rank($\{l_i|H\}$) # ranking of all $\{l_i\}$ that are in H from high to low
    $R \leftarrow$ arg(r_1) # GO-term that belongs to the highest rank
    for arg(r_1) find shortest path(s), p, to root
    delete all nodes in H that are on $p \setminus$ arg(r_1)
    $n = |H|$
R contains the prioritized GO-terms.

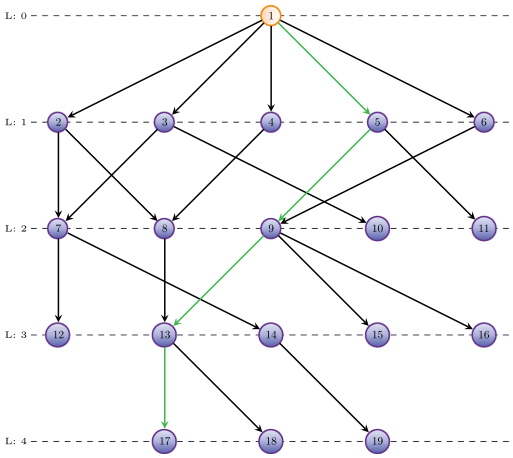


Figure 3. Shown is a path (green) in a GO-DAG, where nodes correspond to GO-terms. Along this path, the biological semantics increases from node to node the further down one traverses the path.

Our algorithm applies the above described logic iteratively, by starting from the GO-term at the highest level and searches all shortest paths to the root node. Then all GO-terms along these shortest paths are removed from the list and the procedure starts over. See Fig. 3 for a visualization. In this figure, one shortest path from node 17 to the root node is shown. The pseudo-code of this is shown in Algorithm 3. Here XX corresponds to BP, MF or CC. The algorithm guarantees that for a non-empty list, H, of GO-terms the resulting set, R, containing the prioritized GO-terms consists of at least one GO-term. For instance, say $H = \{5, 9, 17\}$. Then our algorithm starts at node 17 and searches all shortest paths to the root. One of these is highlighted in green in Fig. 3. As a result, the nodes 5 and 9 are eliminated because they appear on a lower hierarchy level than node 17. In this case, the final result of our algorithm gives $R = \{17\}$.

Overall, our prioritizing algorithm provides a parameter- and assumption-free, non-redundant ranking of GO-terms that exploits only vertical structural information of GO.

Technical details about GO. For the construction of the various DAGs, we are only utilizing information from GO-basic. The information about this can be obtained from the go-basic.obo file, which can be obtained from the Gene Ontology website (http://geneontology.org/docs/download-ontology/). This file contains the basic version of GO and it is guaranteed that the resulting DAG is acyclic and annotations can be propagated through the graph. We would like to note that the relations included in this, i.e., "is_a", "part_of", regulates, "negatively_regulates" and "positively_regulates" also guarantee transitivity (NB: transitivity is not obeyed by "has_part" relations which are included in GO-core available from the *go.obo* file via the GeneOntology website).

Results

In the following sections, we highlight some of the features provided by the `GOxploreR` package and show some example applications.

Structural exploration of GO. In Table 1, we show an overview of the organisms supported by the `GOxploreR` package. Overall, at the moment ten organisms are supported corresponding also to the main organisms within the GO database. The second column in Table 1 shows the option name as used for arguments in functions.

For instance, the following command gives for the gene list 'c(10212, 9833)' containing Entrezgene IDs information about the associated GO-terms and hierarchy levels.

Organism	Option name	Genes	Levels	BP-terms
Human	"*Homo sapiens*"/"Human"	19155	19	12436
Mouse	"*Mus musculus*"/"Mouse"	20929	18	12328
Caenorhabditis elegans	"*Caenorhabditis elegans*"/"Worm"	14697	17	3689
Drosophila melanogaster	"*Drosophila melanogaster*"/"Fruit fly"	12683	18	5323
Rat	"*Rattus norvegicus*"/"Rat"	19383	18	11584
Baker's yeast	"*Saccharomyces cerevisiae*"/"Yeast"	5502	17	3050
Zebrafish	"*Danio rerio*"/"Zebrafish"	20718	18	5404
Arabidopsis thaliana	"*Arabidopsis thaliana*"/"Cress"	25891	17	4059
S. pombe	"*Schizosaccharomyces pombe*"/"Fission yeast"	5055	16	2973
Escherichia coli	"*Escherichia coli*"/"*E.coli*"	3449	15	1491

Table 1. An overview of the organisms supported by the `GOxploreR` package.

```
> Gene2GOTermAndLevel(genes = c(10212, 9833), organism = "Human", domain = "BP")
   entrezgene id       goid ont. level
1          10212 GO:0006397   BP     8
2          10212 GO:0008380   BP     8
3          10212 GO:0006406   BP    12
4          10212 GO:0000398   BP    11
5          10212 GO:0006405   BP     8
6          10212 GO:0031124   BP     9
7           9833 GO:0006468   BP     7
8           9833 GO:0016310   BP     5
9           9833 GO:0018108   BP     9
10          9833 GO:0007049   BP     2
11          9833 GO:0006915   BP     4
12          9833 GO:0035556   BP     5
13          9833 GO:0008283   BP     1
14          9833 GO:0043065   BP     7
15          9833 GO:0046777   BP     8
16          9833 GO:0030097   BP     7
17          9833 GO:0000086   BP     6
18          9833 GO:0061351   BP     2
19          9833 GO:0008631   BP     7
```

In case a list of GO-terms is already available the corresponding hierarchy levels can be obtained with the command 'GOTermXXOnLevel'. Here 'XX' is either BP, MF or CC. In the following, 'XX' corresponds always to one of these three domains.

```
> goterms <- c("GO:0009083","GO:0006631","GO:0006629","GO:0014811","GO:0021961")
> GOTermBPOnLevel(goterm = goterms)
         Term Level
1 GO:0009083     8
2 GO:0006631     7
3 GO:0006629     3
4 GO:0014811    19
5 GO:0021961    15
```

For the analysis of enriched GO-terms, one frequently wants to limit such an analysis to more informative GO-terms which are located toward higher hierarchy levels. In order to obtain all GO-terms located on a specific hierarchy level one can use the function 'Level2GOTermXX'.

```
> Level2GOTermBP(level = 17, organism = "Human")
 [1] "GO:2000321" "GO:0010880" "GO:2000320" "GO:0045630" "GO:2000703"
 [6] "GO:2000734" "GO:0031587" "GO:0045627" "GO:0045629" "GO:0045626"
[11] "GO:0021808" "GO:0060315" "GO:0060316" "GO:0021836" "GO:0021972"
[16] "GO:0031586" "GO:0021817" "GO:0097379" "GO:0021816" "GO:0097380"
```

It is interesting to highlight that the children of a GO-term in a GO-DAG can 'jump' to different levels. For instance, using the function 'GOTermXX2ChildLevel' gives the GO-terms as well as the corresponding hierarchy levels of these.

```
> GOTermBP2ChildLevel(goterm = "GO:0007635")
$Terms
[1] "GO:0007636" "GO:0007637" "GO:0042048" "GO:0061366"

$Level
[1] 5 7 4 6
```

Here the GO-term "GO:0007635" is on level 3, however, its children are not only on level 4. The reason for this is that in GO there are no cross links on the same level. That means the children of any GO-term are always on a lower level because the terms are more specific. This implies that "GO:0007636" which is located on level 5 has (at least one) parent node located on level 4. In order to find this parent(s) we can use the following.

```
go <- Level2GOTermBP(level = 4)
L <- length(go)
go.par <- c()
for(i in 1:L){
        go.ch <- GOTermBP2ChildLevel(goterm = go[i])$Terms
        if( length(which(go.ch == "GO:0007636")) ){
                go.par <- c(go.par, go[i])
        }
}
```

In this case there are 1166 GO-terms on level 4 and the only parent of "GO:0007636" is "GO:0007630".

It is important to note that GO does not only provide one DAG but several different ones. The reason for this is that each organism has a specific number of genes, and from these genes one obtains only a subset of all GO-terms that are connected to an organism. In total there are eleven GO-DAGs available from `GOxploreR`, ten for the organisms and one for all GO-terms.

In order to demonstrate the differences in the GO-terms for different organisms, we show in Fig. 4 the distribution of GO-terms of BP for human (top), zebrafish (middle) and *E. coli* (bottom). The x-axis corresponds to the hierarchy level of the corresponding GO-DAG of BP. As one can see for human one has a GO-DAG with 19 hierarchy levels whereas for zebrafish one has 16 and for *E. coli* 14. Furthermore, also the number of GO-terms on these levels is considerably different from each other as can be seen from the counts (number of GO-terms) on the y-axis. In Table 1, we show an overview of the number of levels (column four) and the number of GO-terms of BP (column five) for all ten organisms. For completeness, we want to mention that if one does not specify the organism in the command 'Level2GOTermBP' one can obtain a total number of 29698 GO-terms of BP for all levels.

Structure-based ranking of GO-terms. Maybe the most popular application of GO is the identification of enriched GO-terms for a list of genes. Unfortunately, as a result from such an analysis it is not uncommon to find large numbers of GO-terms making a focused discussion very difficult. However, a GO-DAG provides information that can be utilized for an exploratory analysis of such a list. Specifically, the hierarchy levels of GO-terms can be utilized. Despite the fact that a GO-level is not an absolute indicator for biological specificity it provides still valuable information[26]. Using our function *GOTermBPOnLevel* gives the GO-levels of BP for a list of GO-terms allowing, e.g., a simple ordering for complementing an enrichment analysis.

For instance, in Fig. 5A, we show results for a list of enriched GO-terms of BP found from an analysis of the breast cancer gene regulatory network[27]. Specifically, the hierarchy levels (x-axis) of these GO-terms (y-axis) are shown in purple. For reasons of comparison, the maximal depth of paths in the GO-DAG passing through these GO-terms is shown in red. As one can see, in all cases, the GO-terms are not at the end of these paths but somewhere situated along the way toward the highest possible (maximal) level that can be reached by passing through the corresponding GO-terms. This information is important because on one-hand one wants to interrogate GO-terms that are biologically specific, i.e., are situated toward the highest hierarchy level of the GO-DAG - for human this would be level 19. On the other-hand not every GO-term is connected to the highest level, i.e., there is no path that would allow to reach the maximal level. Hence, there is a trade-off between absolute and

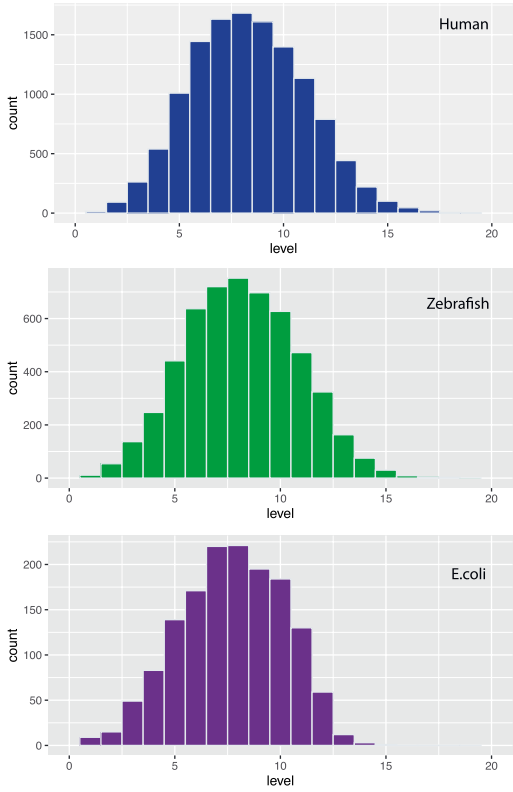


Figure 4. Distribution of GO-terms of BP for human (top), zebrafish (middle) and *E. coli* (bottom). The x-axis corresponds to the hierarchy level of the corresponding GO-DAG.

relative position of a GO-term within a GO-DAG. For this reason, the GO-terms in Fig. 5A are ranked according to the distance between the two points (purple and red).

This trade-off can be formally quantified by the following score,

$$s_t = \text{score} = \frac{\text{level}\,(GO)}{\text{level}_{max}(GO)} \times \frac{\text{level}\,(GO)}{\text{level}_{GO-DAG}(GO)} = p_1(\text{ max path})p_2(GO - DAG). \tag{19}$$

Since the left-hand-side of Eq. (19), i.e, $\frac{\text{level}\,(GO)}{\text{level}_{max}(GO)} \in (0, 1]$, as well as the right-hand-side, i.e., $\frac{\text{level}\,(GO)}{\text{level}_{GO-DAG}(GO)} \in (0, 1]$ the resulting score is also positive and at most one. Hence, the score, s_t, is a product of two probabilities, i.e., $s_t = p_1(\text{ max path})p_2(GO - DAG)$ allowing to optimize the trade-off between both objectives.

The resulting score s_t is shown in Fig. 5B. As one can see, the ranking of GO-terms is similar to Fig. 5A but not identical because Fig. 5A considers for the ranking only the relative distance between the actual and the maximal attainable position in a GO-DAG. Hence, both figures provide slightly complementary information. For our example GO:0006614 (SRP-dependent cotranslational protein targeting to membrane) and GO:0006613 (cotranslational protein targeting to membrane) have the highest score, which are interestingly directly connected in the GO-DAG. Overall, in general this information enables an exploratory analysis of GO-terms which complement the obtained p-values from an enrichment analysis.

In `GOxploreR`, such an analysis can be performed by using the commands *distRankingGO* and *scoreRankingGO*, i.e., the results in Fig. 5A,B can be obtained by

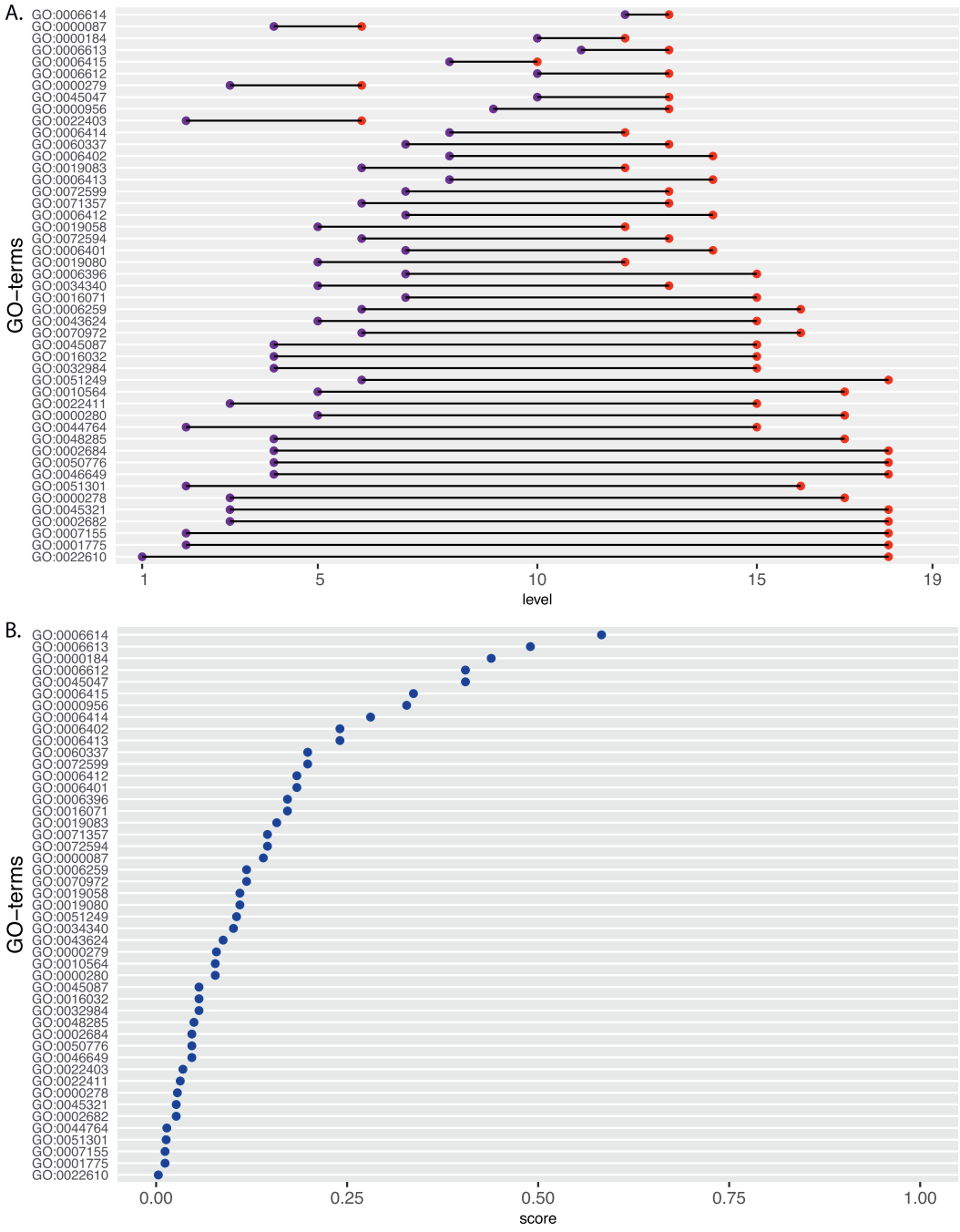


Figure 5. (**A**) The hierarchy levels for a list of GO-terms (y-axis) are shown in purple and the hierarchy levels for the maximal depth of paths in the GO-DAG passing through these GO-terms is shown in red. (**B**) Rank ordered GO-terms according to the score s_t.

```
distRankingGO(goterm = Terms, domain = "BP", plot = TRUE)
scoreRankingGO(goterm = Terms, domain = "BP", plot = FALSE)
```

Reduced GO-DAG. The starting point for many different types of analyses is usually a visualization of the data in order to derive an intuition about the information contained in the data. Unfortunately, for unfiltered GO-terms such a visualization is not feasible because the entire GO-DAG of an organism is too large containing thousands or even tens of thousands of GO-terms (see Table 1). For instance, even the smallest organism with respect to GO-terms of BP consists of 1491 nodes in the corresponding GO-DAG, distributed over 15 hierarchy levels. A graph of such a size cannot be visualized in an insightful way[28]. For this reason, we introduce a so called *reduced GO-DAG* that allows an easy visualization.

The underlying idea of such a reduced GO-DAG is a mapping from GO-terms into three node categories, namely: regular nodes (RN), jump nodes (JN) and leaf nodes (LN). A GO-term is called a 'regular node' (RN) if all its children are on the next level, a GO-term is a 'jump node' (JN) if it has children and at least one of these is not on the next level and a GO-term is a 'leaf node' (LN) if it has no children at all. Such a mapping is obtain by the function *getGOcategory*.

As an example, Fig. 6A shows the reduced GO-DAG of MF for *C. elegans*. This GO-DAG contains only 37 category nodes, i.e., RNs, JNs or LNs, which summarize all 2102 GO-terms of MF for this organism on 14 hierarchy levels. That means only category nodes are shown that contain at least one GO-term, allowing a system-wide view of all MFs of *C. elegans*. Importantly, a reduced GO-DAG has the same number of hierarchy levels as the original GO-DAG because the mapping into category nodes does not effect the hierarchy levels. This holds for all GO-DAG. The following code demonstrates how the information shown in Fig. 6A can be obtained.

```
visRDAGMF(organism = "Caenorhabditis elegans", plot = TRUE)
```

Similar visualizations are possible for all other organisms because even for human, there are only 52 (BP), 38 (MF), 43 (CC) nodes in the resulting reduced GO-DAG for the corresponding domains.

In case one has a list of GO-terms, one can also perform such a mapping only for this limited number of GO-terms. Furthermore, also a visualization for this sub-set of all GO-terms can be obtained using the function *visRDAGMF*. Overall, a reduced GO-DAG helps in simplifying the complexity provided by the gene ontology especially with respect to the connectivity between the GO-terms. This enables a general visualization for an exploratory analysis of system-wide information propagation capabilities.

Prioritizing GO-terms. Finally, `GOxploreR` provides a function called *prioritizedGOTerms* for prioritizing GO-terms. The idea is to go beyond the ordering of GO-terms for a provided list of GO-terms to eliminate selected terms that are capturing redundant and less biologically specific information; see the discussion of Fig. 6B below.

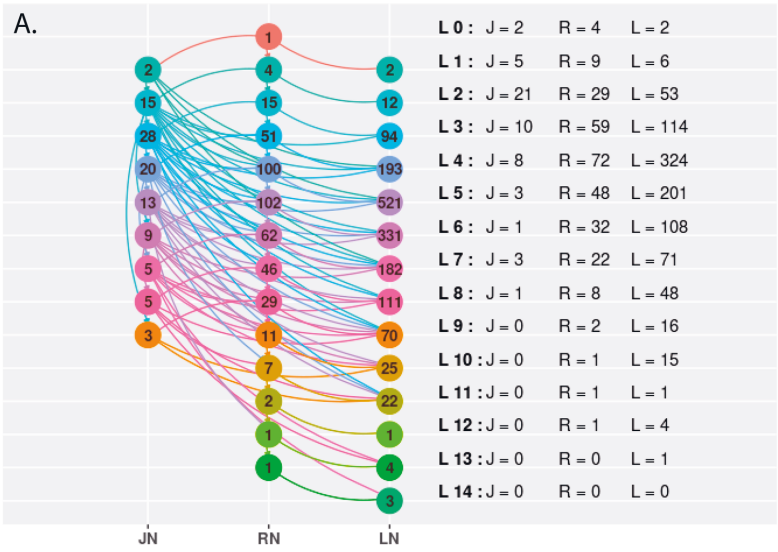
In order to realize an implementation for such a function, we apply the following strategy (see Methods Sec. 2.4 for technical details). Specifically, it is known that the comparison of GO-terms with respect to their biological meaning is complex. However, the comparison of GO-terms that can be found along a path is much simpler because the higher a level of a GO-term, the more specific is its biological information[26]. That means traversing a path *vertically* toward higher levels increases the biological specificity of GO-terms implying that the GO-term at the end of a path is the most interesting one. Hence, by eliminating all GO-terms that are together on a path, except the one on the highest level, results in a prioritizing of terms with respect to the semantic meaning of GO-terms. The function *prioritizedGOTerms* implements this strategy. In Fig. 6B, we show visualized of this. Here one path is highlighted containing three GO-terms (GO:1, GO:2, and GO:3) whereas GO:3 has the highest level. This results in an elimination of GO:1 and GO:2. Similarly, all other paths are explored resulting in GO:1 and GO:6 as output of the prioritizing algorithm.

As an example, we investigate a list of GO-terms that was obtained from analyzing a gene regulatory network of *S. cerevisiae*[29]. The original list contains 30 different GO-terms of BP[29], each significantly enriched with a significant p-value. Application of our function *prioritizedGOTerms* for prioritizing GO-terms results in only 5 GO-terms, shown in Table 2. Each of these 5 GO-terms is located on a separate brunch of the underlying GO-DAG between which no paths exist. Hence, despite of a certain similarity of the biological processes, e.g., for metabolic or mitochondrial processes, each of these terms is from a different, separate semantic category because otherwise connections with the DAG would exist. Such an analysis complements available p-values and gives further information on which GO-terms a follow-up analysis could focus on.

Overall, the function *prioritizedGOTerms* can prioritize a list of GO-terms with information about the semantic information content of a GO-DAG as provided by the level of GO-terms. If desired, a separate visualization could be obtained only for these GO-terms by using the function *visRDAGsubMF*.

Discussion

In this paper, we introduced the `R` package `GOxploreR` and highlighted some of the functionality it provides. Overall, `GOxploreR` provides functions and algorithms for four different types of analyses. Specifically, `GOxploreR` enables a (1) direct access to structural features of GO, (2) structure-based ranking of GO-terms, (3) mapping to a reduced GO-DAG and (4) prioritizing of GO-terms.



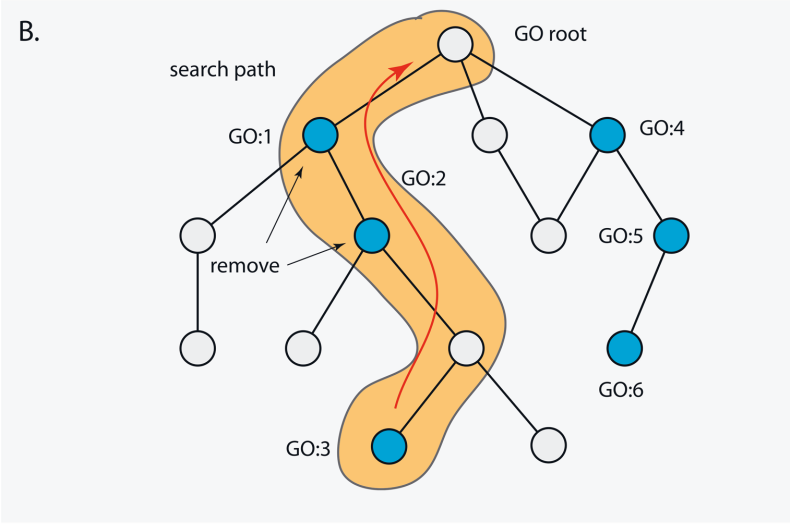


Figure 6. (**A**) Shown is a reduced GO-DAG of MF for *C. elegans*. The whole GO-DAG contains only 37 category nodes, i.e., RN, JN or LN and summarizes all 2103 GO-terms of MF for this organism. (**B**) Underlying idea for prioritizing GO-terms in a general DAG. Shown is one search path. Nodes in blue correspond to GO-terms in a given list.

The first three features of `GOxploreR` permit an exploratory analysis of GO-terms and GO-DAGs whereas the fourth feature provides a dedicated algorithm for a particular problem. Despite the fact that it is well-known that GO has the structure of a DAG, there are surprisingly few tools allowing a direct assess to structural, i.e., graph-based information of GO. Hence, our features and the corresponding functions help in utilizing this rich source of information which is in our opinion so far largely underexplored. A reason for this lack could be that the conceptual realization and implementation of graph-based algorithms is not straight forward requiring inter- and transdisciplinary knowledge of graphs and the underlying biology.

GO-term	GO-level	Description	*p* value	# genes
GO:0006364	9	rRNA processing	1.6e−39	237
GO:0032543	8	Mitochondrial translation	4.2e−167	100
GO:0044257	6	Cellular protein catabolic process	2.0e−78	347
GO:0019752	5	Carboxylic acid metabolic process	3.5e−67	370
GO:0007005	4	Mitochondrion organization	3.0e−168	282

Table 2. Using GOxploreR one can prioritize lists of GO-terms. The table shows results for significant GO-terms from analyzing a gene regulatory network of *S. cerevisiae*[29] after the application of our prioritizing algorithm. The GO-terms are for BP and complement p-values obtained from an independent enrichment analysis.

One important novelty of `GOxploreR` is to provide a mapping from a GO-DAG to a reduce GO-DAG. This leads to a tremendous reduction in complexity of graphs because a GO-DAG can contain thousands of nodes, depending on the organism and the domain, i.e., BP, MF or CC. In contrast, a reduced GO-DAG has at most three nodes of the categories, JN (jump node), RN (regular node) or LN (leaf node) on each hierarchy level. The idea behind this mapping is inspired by the detection of differentially expressed genes (DEG)[30]. While the expression level of a gene is continuous, a DEG analysis performs a kind of classification of the expression level into two categories: active and inactive. This allows a reduction in the complexity of the gene expression level by capturing simplified yet essential information. Our mapping from a GO-DAG to a reduce GO-DAG follows a similar strategy by capturing simplified yet essential information of the connection between GO-terms. As far as we know, `GOxploreR` is the only package that provides such a mapping and reduction in the GO complexity.

Another novelty of the `GOxploreR` package is to provide visualizations of reduced GO-DAGs. This feature is directly enabled by the tremendous reduction in complexity of the mapping from a GO-DAG to a reduce GO-DAG because the visualization of a DAG containing thousands of nodes (see Table 1) is not feasible. In contrast, a reduce GO-DAG permits such a visualization allowing to obtain an overview of the biological information processing of the entire ontology. Given the novelty of a mapping from a GO-DAG to a reduce GO-DAG other packages that provide also visualization capabilities do not offer this particular visualization.

Finally, the `GOxploreR` package provides a prioritizing algorithm. The idea of this algorithm is to go beyond the ordering of GO-terms for a given list of GO-terms, and to eliminate GO-terms capturing redundant biological information. For the prioritizing of GO-terms in a list, we utilized the fact that the higher a level of a GO-term is the more specific is its biological information[26]. That means *vertically* one wants to traverse a DAG as far down as possible because the end of a path is most specific compared to all other GO-terms along this path. Our algorithm applies this logic iteratively by starting from the GO-term at the highest level and searches all (shortest) paths to the root node. Then all GO-terms along these shortest paths are removed from the list and the procedure starts over; see Fig. 1 for a visualization. As a result, one obtains a prioritizing of GO-terms that is a parameter- and assumption-free algorithm which removes redundant GO-terms by exploiting only vertical structural information of a GO-DAG. Hence, the output of our prioritizing algorithm is a non-redundant ranking of GO-terms.

We would like to highlight that there is a crucial difference between our prioritizing algorithm and approaches based on the semantic similarity of genes[31,32]. The difference is that we utilize only vertical information from a GO-DAG. This implies that there is no need for comparing GO-terms horizontally because they cannot be connected by any path (besides over the root node). However, this horizontal comparison is usually the problem since the biological significance of different GO-terms on the same hierarchy level can be different. This simplifies the analysis yet allows the elimination of redundant GO-terms. The resulting list of GO-terms maybe be further reduced, however, not without making additional assumptions, e.g., in the form of semantic similarity measures. A common problem with the latter is that there is not one but many different measures for semantic similarity all of which are non-trivial in their definition and interpretation[33]. In contrast, our prioritizing algorithm is parameter- and assumption-free allowing to remove redundant GO-terms by exploiting only vertical structural information along paths of a GO-DAG. Another fundamental difference between our prioritizing algorithm and semantic similarity measures is that our algorithm focuses on GO-terms and not on genes. This facilitates a general systems view on the underlying problem from which the GO-terms have been obtained as represented by systems biology[34,35].

In Table 3, we compare the capabilities of the `GOxploreR` package with other software tools available for analyzing GO. The first column shows the name of the software whereas the remaining columns refer to various features. Specifically, the second column indicates if a software tool is available as an R package and the third column refers to direct assess of structural information provided by a GO-DAG. Examples thereof are the hierarchical level of a GO-term, the GO-terms on a certain hierarchy level or the adjacency matrix of a DAG. The fourth column is about identifying the enrichment of GO terms, whereas the fifth column is about the availability of reduced GO-DAGs and the sixth column refers to a prioritizing algorithm for a list of GO-terms.

As one can see from Table 3, the `GOxploreR` package is considerably different from all the other software tools, hence, providing novel and complementary analyses functionality. Importantly, `GOxploreR` is available as R package allowing the easy utilization of it within existing analysis pipelines for their extensions. Hence, `GOxploreR` does not provide dead-end functionality via web-interfaces but enables future biomedical data science projects[36].

Name	R Package	Direct structural information to GO	Enrichment	Reduced GO-DAG	Prioritizing GO-terms
GOxploreR	Yes	Yes	No	Yes	Yes
OntologyTraverser[37]	Yes	Partly	Yes	No	No
Categorizer[38]	No	No	Yes	No	No
G-SESAME[39]	No	No	No	No	No
GOrilla[40]	No	No	Yes	No	No
GOGrapher[41]	No	Partly	No	No	No
agriGO[42]	No	No	Yes	No	No
topGO[43]	Yes	Yes	Yes	No	No
GOdb[44]	Yes	Yes	No	No	No

Table 3. A comparison of the capabilities of various software tools for analyzing GO.

Conclusion

In this paper, we introduced the `R` package `GOxploreR`, available from CRAN (after acceptance of the paper). `GOxploreR` is a versatile tool that can be applied to any list of GO-terms from an upstream analysis as a result from studying, e.g., differentially expressed genes, GWAS, biomarkers, gene sets or gene regulatory network studies[7–13]. Its main features include:

1. A direct access to structural features of GO.
2. A structure-based ranking of GO-terms.
3. A mapping from a GO-DAG to a reduced GO-DAG.
4. A visualization of reducuded GO-DAGs.
5. An algorithm for prioritizing GO-terms.

Given the lack of tools for exploring the DAG-structure of GO from a graph theoretical perspective, `GOxploreR` complements non-structural analysis tools. Overall, `GOxploreR` has the potential to enhance studies investigating differentially expressed genes, GWAS (genome-wide association study), biomarkers, gene sets or gene regulatory network studies significantly because the obtained information has a clear interpretation directly derived from the gene ontology knowledge base and is not based on additional assumptions.

Received: 20 January 2020; Accepted: 17 August 2020
Published online: 07 October 2020

References

1. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Gene Ontol. Consort. Nat. Genet.* **25**, 25–29 (2000).
2. Consortium, G. O. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2018).
3. Tweedie, S. *et al.* Flybase: enhancing drosophila gene ontology annotations. *Nucleic Acids Res.* **37**, D555–D559 (2008).
4. Boyle, E. I. *et al.* GO::TermFinder-open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004).
5. Binns, D. *et al.* Quickgo: a web-based tool for gene ontology searching. *Bioinformatics* **25**, 3045–3046 (2009).
6. Jacobson, M., Sedeño-Cortés, A. E. & Pavlidis, P. Monitoring changes in the gene ontology and their impact on genomic data analysis. *GigaScience* **7**, giy103 (2018).
7. Young, M., Wakefield, M., Smyth, G. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
8. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
9. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).
10. Arciero, C. *et al.* Functional relationship and gene ontology classification of breast cancer biomarkers. *Int. J. Biol. Markers* **18**, 241–272 (2003).
11. Mooney, M. A., Nigg, J. T., McWeeney, S. K. & Wilmot, B. Functional and genomic context in pathway analysis of GWAS data. *Trends Genet.* **30**, 390–400 (2014).
12. Schaid, D. J. *et al.* Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet. Epidemiol.* **36**, 3–16 (2012).
13. Cun, Y. & Fröhlich, H. Biomarker gene signature discovery integrating network knowledge. *Biology* **1**, 5–17 (2012).
14. Hoehndorf, R., Schofield, P. N. & Gkoutos, G. V. The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.* **16**, 1069–1080 (2015).
15. Ten Blake, J. A. Quick tips for using the gene ontology. *PLoS Comput. Biol.* **9**, e1003343 (2013).
16. Emmert-Streib, F. & Dehmer, M. Networks for systems biology: conceptual connection of data and function. *IET Syst. Biol.* **5**, 185 (2011).
17. Aittokallio, T. & Schwikowski, B. Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.* **7**, 243–255 (2006).
18. Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288–289 (2008).
19. Martin, D. *et al.* GOToolBox: functional analysis of gene datasets based on gene ontology. *Genome Biol.* **5**, R101 (2004).
20. Ye, J. *et al.* WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* **34**, W293–W297 (2006).
21. Beißbarth, T. & Speed, T. P. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* **20**, 1464–1465 (2004).

22. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2006).
23. du Plessis, L., Škunca, N. & Dessimoz, C. The what, where, how and why of gene ontology? A primer for bioinformaticians. *Brief. Bioinform.* **12**, 723–735 (2011).
24. Grossmann, S., Bauer, S., Robinson, P. N. & Vingron, M. Improved detection of overrepresentation of gene-ontology annotations with parent-child analysis. *Bioinformatics* **23**, 3024–3031 (2007).
25. Mazandu, G. K. & Mulder, N. J. Information content-based gene ontology functional similarity measures: Which one to use for a given biological data type?. *PLoS ONE* **9**, e113859 (2014).
26. Dennis, G. *et al.* DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, R60 (2003).
27. Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B. & Dehmer, M. The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Front. Genet.* **5**, 15 (2014).
28. Tripathi, S., Dehmer, M. & Emmert-Streib, F. NetBioV: an R package for visualizing large-scale data in network biology. *Bioinformatics* **30**, 2834–2836 (2014).
29. de Matos Simoes, R. & Emmert-Streib, F. Bagging statistical network inference from large-scale gene expression data. *PLoS ONE* **7**, e33624 (2012).
30. Dudoit, S., Yang, Y. H., Callow, M. J. & Speed, T. P. Statistical methods for identifying differentially expressed genes in replicated CDNA microarray experiments. *Statistica Sinica* **12**, 111–139 (2002).
31. Gan, M., Dou, X. & Jiang, R. From ontology to semantic similarity: calculation of ontology-based semantic similarity. *Sci. World J.*https://doi.org/10.1155/2013/793091 *(2013).*
32. Pesquita, C., Faria, D., Falcao, A. O., Lord, P. & Couto, F. M. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* **5**, e1000443 (2009).
33. Pesquita, C. Semantic similarity in the gene ontology. In *The Gene Ontology Handbook* 161–173 (Humana Press, New York, 2017).
34. Emmert-Streib, F. & Glazko, G. Network biology: a direct approach to study biological function. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **3**, 379–391 (2011).
35. Vidal, M. A unifying view of 21st century systems biology. *FEBS Lett.* **583**, 3891–3894 (2009).
36. Emmert-Streib, F. & Dehmer, M. Defining data science by a data-driven quantification of the community. *Mach. Learn. Knowl. Extraction* **1**, 235–251 (2019).
37. Young, A., Whitehouse, N., Cho, J. & Shaw, C. OntologyTraverser: an R package for GO analysis. *Bioinformatics* **21**, 275–276 (2004).
38. Na, D., Son, H. & Gsponer, J. Categorizer: a tool to categorize genes into user-defined biological groups based on semantic similarity. *BMC Genomics* **15**, 1091 (2014).
39. Du, Z., Li, L., Chen, C.-F., Yu, P. S. & Wang, J. Z. G-sesame: web tools for go-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Res.* **37**, W345–W349 (2009).
40. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinform.* **10**, 48 (2009).
41. Muller, B., Richards, A. J., Jin, B. & Lu, X. Gographer: a python library for go graph representation and analysis. *BMC Res. Notes* **2**, 122 (2009).
42. Tian, T. *et al.* agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129 (2017).
43. Alexa, A. & Rahnenfuhrer, J. topgo: enrichment analysis for gene ontology. R package version 2 (2010).
44. Carlson, M. Go. db: A set of annotation maps describing the entire gene ontology (2016).

Acknowledgements
Matthias Dehmer thanks the Austrian Science Funds for supporting this work (project P30031).

Author contributions
F.E.S conceived the study. K.M., S.T. and F.E.S. conducted the analysis. K.M., S.T., O.Y.H., M.D. and F.E.S. interpreted the results. All authors wrote the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-73326-3.

Correspondence and requests for materials should be addressed to F.E.-S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PUBLICATION

II

Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning

K. Manjang, S. Tripathi, O. Yli-Harja, M. Dehmer, G. Glazko and F. Emmert-Streib

Scientific Reports 11.1 (2021), 1–18
DOI: 10.1038/s41598-020-79375-y

scientific reports

Check for updates

OPEN

Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning

Kalifa Manjang[1], Shailesh Tripathi[1], Olli Yli-Harja[2,3,7], Matthias Dehmer[4,5,6], Galina Glazko[7] & Frank Emmert-Streib[1,8]✉

The identification of prognostic biomarkers for predicting cancer progression is an important problem for two reasons. First, such biomarkers find practical application in a clinical context for the treatment of patients. Second, interrogation of the biomarkers themselves is assumed to lead to novel insights of disease mechanisms and the underlying molecular processes that cause the pathological behavior. For breast cancer, many signatures based on gene expression values have been reported to be associated with overall survival. Consequently, such signatures have been used for suggesting biological explanations of breast cancer and drug mechanisms. In this paper, we demonstrate for a large number of breast cancer signatures that such an implication is not justified. Our approach eliminates systematically all traces of biological meaning of signature genes and shows that among the remaining genes, surrogate gene sets can be formed with indistinguishable prognostic prediction capabilities and opposite biological meaning. Hence, our results demonstrate that none of the studied signatures has a sensible biological interpretation or meaning with respect to disease etiology. Overall, this shows that prognostic signatures are black-box models with sensible predictions of breast cancer outcome but no value for revealing causal connections. Furthermore, we show that the number of such surrogate gene sets is not small but very large.

Since the inception of high-throughput technologies the goal has been to utilize such experimental devices not only for obtaining a better elucidation of biology but to translate this knowledge into the clinical practice[1,2]. One particular example for such an application are prognostic studies based on gene expression data[3–5]. In general, the goal of such studies is to select a, preferably small, number of genes as features, called a signature, and to utilize these for predicting the course of a disease or outcome of patients represented by gene expression profiles. The prognostic value of such predictions is quantitatively assessed via a survival analysis allowing to perform a statistical test for detecting differences in different patient groups with respect to 'time to event' information. Due to the generality of 'event', which cannot only be death but also relapse or development of metastasis or organ rejection, prognostic studies are relevant for nearly all patient-related medical investigations. Due to the importance of prognostic studies for clinical applications and their general complexity, statistical aspects of this problem have attracted much attention in the literature. For instance, in[6] the authors addressed the stability of the selection of prognostic predictors for various cancer types. They found that the size of the training data and the patients in it has a crucial effect on this. The same problem has been studied for breast cancer in[7] and the authors found that thousands of patient samples are needed for achieving an overlap of 50% between two predictive sets of genes. Such problems have been confirmed in many comparative investigations of feature selection mechanisms, see, e.g.,[8–10].

[1]Predictive Society and Data Analytics Lab, Tampere University, Tampere, Korkeakoulunkatu 10, 33720 Tampere, Finland. [2]Computational Systems Biology, Tampere University, Tampere, Korkeakoulunkatu 10, 33720 Tampere, Finland. [3]Institute for Systems Biology, Seattle, WA, USA. [4]Steyr School of Management, University of Applied Sciences Upper Austria, 4400 Steyr Campus, Wels, Austria. [5]College of Artificial Intelligence, Nankai University, Tianjin 300350, China. [6]Department of Biomedical Computer Science and Mechatronics, UMIT-The Health and Life Science University, 6060 Hall in Tyrol, Innsbruck, Austria. [7]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, USA. [8]Institute of Biosciences and Medical Technology, Tampere University, Tampere, Korkeakoulunkatu 10, 33720 Tampere, Finland. ✉email: frank.emmert-streib@tuni.fi

For breast cancer, an early study of a prognostic gene expression signature is from[11]. The authors used a 70-gene signature to distinguish good prognosis from bad prognosis groups of patients with stage I or II breast cancer. The outcome of this influential paper sparked many follow-up investigations. For instance, in[12] a 76-gene signature was used predicting development of distant metastases within 5 years of lymph-node-negative primary breast cancer or in[13] an invasiveness signature of a 186-gene signature was used for predicting overall and metastasis-free survival. It is important to note that for all such studies not only the predictive outcome is of value but also the interpretational biological meaning of the used signatures[14]. Specifically, it has been stated in[7] that "A reliable set of predictive genes also will contribute to a better understanding of the biological mechanism of metastasis". This assumption is not limited to the above problem but widely believed to be true in the genomics and translational medicine community. The main purpose of this paper is to refute this assumption.

Our study is different from the above mentioned ones with respect to the following aspects. First, we do not introduce a new procedure for selecting signature genes. Instead, we provide an analysis of previously introduced signatures with respect to their biological meaning. Second, we do not introduce a new validation method because all studied signatures have been previously validated, although we are using an independent validation data set for our study. Third, we do not aim to improve the quality of different prognostic signatures, although we utilize a more stringent statistical assessment, including conservative multiple testing corrections, compared to previous studies. Fourth, we do also not establish a connection between a prognostic signature and disease etiology shedding light on the underlying molecular and cell biological mechanisms. Instead, we investigate the prognostic benefit of random gene sets having a constrained biological meaning. The main purpose of this paper is to systematically demonstrate that sensible prognostic signatures of breast cancer outcome do not have a sensible biological meaning with respect to disease etiology. This is accomplished via *constrained-sampling*, a restricted resampling procedure for constructing random gene sets, which we introduce in this paper.

A central aspect of our constrained-sampling analysis is based on the definition of biological meaning of a set of genes. For this, we are using two different commonly utilized approaches. The first is centered around the meaning of individual 'genes' and the second is based on 'biological processes'. For the gene-based definition of biological meaning, we follow a Mendelian-view whereas for the biological process-based definition representing a systems-view[15], we utilize Gene Ontology (GO)[16] and its underlying hierarchically organized GO-terms in the form of a directed acyclic graph (DAG).

This paper is organized as follows. In the next section, we describe the underlying methodology and the used data. Then we present our results and discuss our findings. This paper finishes with concluding remarks.

Methods

In this section, we provide information about the data and methods used for our analysis.

Gene expression data and BM signatures. Our analysis makes use of two sources of data—gene expression data and sets of breast cancer gene signatures from 48 published studies. For the gene expression data we use two different data sets publicly available. The first gene expression dataset (in the following called NKI breast cancer) is accessible from[17] and it contains 295 breast cancer samples from the Netherlands Cancer institute (a.k.a NKI) cohort. The data were generated by[11]. The gene expression dataset consists of 13108 genes and each sample corresponds to one patient. All patients had stage I or II breast cancer. The dataset is complemented by information about the development of metastases which has been used to indicate an 'event' for survival analysis. The second gene expression dataset (in the following called SWE breast cancer) is from Gene Expression Omnibus (GSE96058)[18]. It contains 30865 genes and samples of the subtypes Basal (360), Her2 (348), LumA (1709), LumB (767) and Normal (225). The data were FPKM normalized and log transformed. The 48 biomarker (BM) signatures we use for our analysis were compiled in[17]. The number of genes in each signature varies, but all the biomarkers together contain 8106 genes. For the NKI gene expression data 5350 genes are present and for the SWE data 5060 genes.

Outcome association. For assessing the prognostic value of gene sets, we perform a survival analysis. Specifically, we perform Kaplan Meier estimates of survival curves and compare these with a Mantel–Haenszel test[19]. Hence, each comparison is characterized by a p-value resulting from such a hypothesis test. The categorization of patients is achieved by the PC1 method, described below. This method separates the patients according to specified gene set. This means that the resulting survival analysis is a function of the gene set used to categorize patients. In Fig. 1, we show an overview of the individual steps involved in our analysis. Overall, our analysis consists of three main steps. First, selection/construction of a random gene set, second, classification of patient samples and, third, performing a survival analysis.

In the next two sections, we specify two different gene removal procedures (GRP) for constructing random gene sets. These procedures implement a constrained-sampling for two different views on biology, a Mendelian-view based on genes (GRP 1) and a systems-view based on biological processes (GRP 2 and GRP 2*).

Gene removal procedure 1. For this analysis, we investigate the prediction capabilities of random gene sets, RGS_i, whereas the genes in RGS_i are randomly sampled from the set $G_i' = G \setminus BM_i$. Here G corresponds to the total number of genes in our breast cancer data set and BM_i is the BM signature of study i, for $i \in \{1, \ldots, 48\}$. The number of genes sampled per random signature is the same as in BM_i, i.e., $|RGS_i| = |BM_i|$. We repeat this sampling 1000 times for each study with and without a Bonferroni correction. From numerical analyses we found that increasing the number of repeats does not lead to different results. In total we study 96, 000 random gene sets that have been constructed in this way. Details of this gene removal process are described as follows:

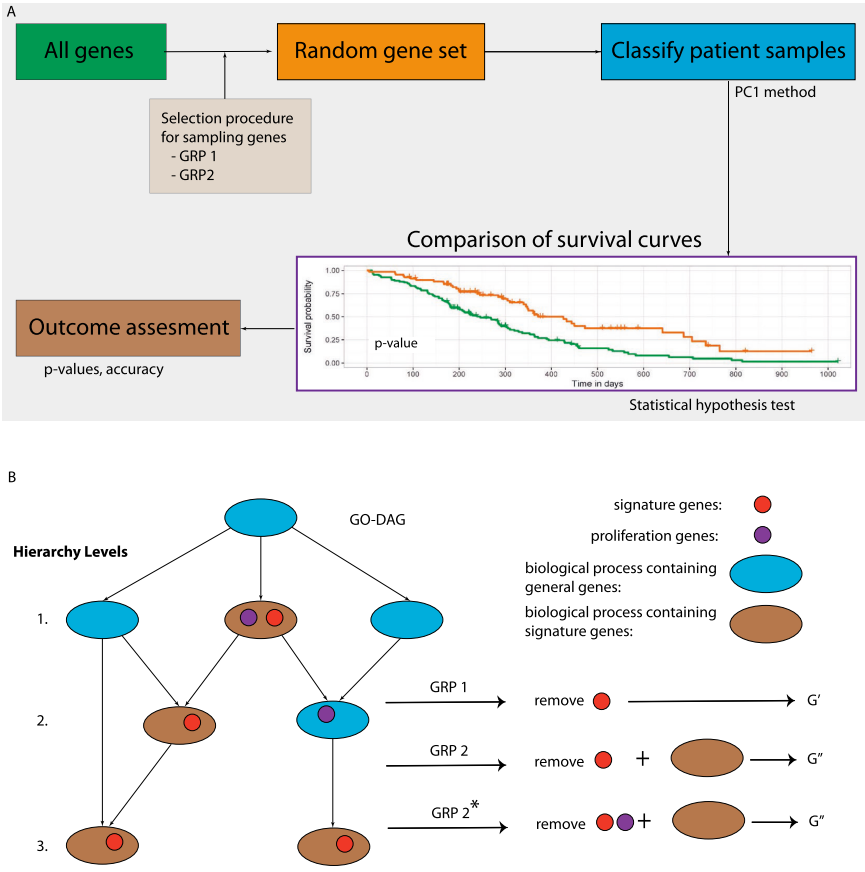


Figure 1. (**A**) Shown is a flowchart of all steps involved in our analysis. (**B**) Visualization of the underlying ideas of GRP 1 and GRP 2. The resulting gene sets G' and G'' are used for sampling random gene sets.

1. G : total number of genes in our breast cancer dataset.
2. $BM_i : \{g_1, \ldots, g_m\}$. BM_i is the gene signature i (i range from 1 to 48) and $g_1, \ldots, g_m$ are the genes in the corresponding signature.
3. For each biomarker set i:

   (a) Removing biomarker genes in signature BM_i from G. This gives a new set of genes G'_i with $G'_i = G \setminus BM_i$.
   (b) From G', we sample new sets of biomarker genes of size $|BM_i|$ and perform the prognostic task. This is repeated 1000 times for each study i.
   (c) Application of a Bonferroni correction to the p-values.
   (d) Assessing the performance for a significance level of α.

Overall, gene removal procedure 1 constructs random gene sets by removal of BM signatures. If a random gene set has a significant p-value, we call it a *surrogate gene set* because it has the same prognostic prediction capabilities as a BM signature and hence it is a surrogate for this.

Gene removal procedure 2. For this analysis, we do not only remove BM signatures, but we remove also genes that belong to the same biological processes as the genes in the BM signatures. Due to the fact that according to the gene ontology (GO) database[16] the biological processes are hierarchically organized, we approach this analysis iteratively by removing successively genes of biological processes on the same hierarchy level[20]. Details of this gene removal process are described as follows:

1. G : total number of genes in our breast cancer dataset.
2. $BM_i : \{g_1, \ldots, g_m\}$. BM_i is the gene signature i (i range from 1 to 48) and $g_1, \ldots, g_m$ are the genes in the corresponding signature.
3. Removing biomarker genes in signature BM_i from G. This gives a new set of genes G_i' with $G_i' = G \setminus BM_i$.
3* Optional step: Removing proliferation genes in PG from G. This gives a new set of genes $G_i'^*$ with $G_i'^* = G_i' \setminus PG$.
4. Mapping of the genes in BM_i to GO-terms and the corresponding hierarchy levels. This gives:

$$BM_i = \{g_1, \ldots, g_m\} \rightarrow \{(GO_1, L_1), \ldots, (GO_t, L_t)\}. \tag{1}$$

Note, each gene can be connected to more than one GO-term. For this reason $m \leq t$.
5. Ranking of the GO-terms in descending order with respect to the hierarchy levels.
6. For each biomarker set i: Loop-over the hierarchy levels l in descending order, i.e., for $l \in \{L_{max}(i), \ldots, L_{min}(i)\}$. Here $L_{max}(i)$ is the highest hierarchy level of biomarker set i and $L_{min}(i)$ is the lowest hierarchy level.

(a) Delete all the genes associated with GO-terms on level l. This results in a new gene set given by $G'' = G' \setminus D$, where D is the set of genes having GO-terms on level l.
(b) From G'', we sample new sets of biomarker genes of size $|BM_i|$ and perform the prognostic task. This is repeated 1000 times for each hierarchy level l.
(c) Application of a Bonferroni correction to the p-values.
(d) Assessing the performance for a significance level of α.
(e) Set $G' = G''$. Stop if $l = L_{min}(i)$ or $|G''| < |BM_i|$.

In the above procedure, the set PG is the gene set consisting of genes related to proliferation. The genes in PG have been defined in[21] and consist of the signature genes of Whitfield[22] and meta-PCNA[17]. In total PG contains 664 genes. Step 3* is an optional step that removes additionally proliferation genes. When step 3* is used, we call the procedure GRP 2*, whereas when step 3* is not used, we call the procedure GRP 2.

Put simply, procedure GRP 2 removes first all biomarker genes (see step 3) and then iteratively removes genes belonging to the same biological processes as the signature genes (see step 6) from the highest hierarchy level L_{max} to the lowest hierarchy level L_{min}. That means at the end a set of genes G'' is obtained that contains neither signature genes nor genes the belong to the same biological processes as the signature genes regardless of the hierarchy level. Results for G'' for intermediate hierarchy levels l contain a certain overlap with biological processes as indicated by l. All sets G'' are treated in a similar way, i.e., the prognostic task is performed and assessed.

We assess the prediction results again by the p-values from the survival analysis. In addition, we determine the accuracy of predictions by declaring significant p-values as true positives (TPs) and non-significant results as false negatives (FNs). This allows the estimation of accuracy values, i.e., $Acc = (TP + TN)/(TP + TN + FP + FN)$ by $Acc = TP/FN$[23]. These evaluations are obtained for each hierarchy level.

Overall, gene removal procedure 2 constructs random gene sets by removal of BM signatures and biological process related genes. Also here a random gene set with a significant p-value is call a *surrogate gene set*. In Fig. 1 B, a visualization of GRP 1, GRP 2 and GRP 2* is shown.

Categorize patient samples. For categorizing the samples of the patients, the PC1 stratification method is used. This method is based on a principal component analysis (PCA). The principal component analysis is a dimensionality reduction technique (this involves reducing the size of the data set). The goal is to transform large data set into smaller ones. This method trades a little accuracy for simplicity, thus achieving interpretability as well as minimal loss of information. Using the "prcomp" function available in R, the first principal component (PC1) of the signature is derived. The patients are then divided into two groups according to the median of the PC1. Specifically, a sample is categorized as group -1 if the PC1 is below the median value and as group $+1$ if the PC1 is above the median value.

For this analysis, a gene expression matrix of the form $X \in \mathbb{R}^m \times \mathbb{R}^n$ is used whereas m is the number of genes and n is the number of samples. Importantly, m corresponds to a particular gene set and not all genes available. Above, we described two different procedures for constructing such gene sets. Other sets we use for our analysis are the BM signatures themselves.

Survival analysis. For assessing the prognostic value of gene sets, we perform a survival analysis. Specifically, we perform Kaplan Meier estimates of survival curves and compare these with a Mantel–Haenszel test[19]. Hence, each comparison is characterized by a p-value resulting from such a hypothesis test. The categorization of patients is achieved by the PC1 method, described above. This method separates the patients according to a specified gene set. Therefore, the resulting survival analysis depends on this gene set.

Definition: biological meaning. In this paper we use the term 'biological meaning' in a well-defined way. This definition is based on gene ontology (GO)[16]. Specifically, the biological meaning of a gene is given by the GO-terms this gene is associated with as provided by GO. Similarly, the biological meaning of a set of genes is provided by the union of the sets of GO-terms of the individual genes.

Results

Our analysis is structured into three main parts. In the first part, we study characteristics of the 48 BM signatures individually and comparatively. In the second and third part, we study prognostic prediction capabilities of random gene sets, systematically constructed with two different procedures.

Biomarker set sizes and GO-term in signatures. In Fig. 2A, we show an overview of the total number of genes in each signature. The name of the signatures are on the y-axis and the x-axis provides information about the size of the BM signatures.

From this figure, one can see that the signature by Adorno and Pei contains the least number of genes (2) whereas Hua has the largest number (1345 genes). That means the size of the signatures varies considerably among the studies and the average size of a signature is 168.9 genes.

In Fig. 2B and C, we show information about associated GO-terms with the genes in the signatures for the categories: Biological process (BP), molecular function (MF), and cellular component (CC). Currently, there are in total 29, 699 GO-terms from BP, 4202 GO-terms from CC and 11, 148 GO-terms from MF. In Fig. 2B, we show the absolute number of GO-terms in each study with respect to BP (green), MF (red) and CC (blue) whereas Figure 2C shows the corresponding percentage with respect to the total number of GO-terms for each category (i.e., BP, MF and CC).

Overall, from Fig. 2 B one can see that the present GO-terms in the signatures is considerably different from each other. This variation is particularly large for GO-terms of BP (green). Interestingly, if one considers the percentage of present GO-terms (see Fig. 2C) then the differences between the three GO categories (i.e., BP, MF and CC) become much smaller, although, also on this scale the differences between studies are considerable. The average number of GO-terms is 996.7 for BP, 277.9 for MF and 204 for CC and the average percentage is 0.034 for BP, 0.025 for MF and 0.049 for CC.

Using a Spearman rank correlation test, we investigate if the order of the size of biomarker sets (see Fig. 2A) is conserved by the number of GO-terms (see Fig. 2B). As a result, we find p-values of $1.311469e-28$ for BP (green), $3.44238e-35$ for MF (red) and $2.96905e-29$ for CC (blue). Due to the fact that the percentage of GO-terms shown in Fig. 2C has the same order as for the number of GO-terms in Fig. 2B, a comparison of these results leads to the exact same p-values. Overall, the above p-values indicate that the order of all comparisons is highly statistically significant for any sensible significance level α. Therefore, the ranking of the biomarker sets with respect to their size is similar to the ranking according to their number of GO-terms, which implies that larger BM signatures contain more GO-terms.

Pairwise similarity of signatures. For our next analysis, we perform a pairwise comparison of the BM signatures. That means, we study the overlap of common genes and GO-terms among different signatures. In Fig. 3A and B, the results from these pairwise comparisons are shown in form of heat maps. Formally, we define the overlap as follows. Let S_i and S_j be two signature sets consisting either of genes or GO-terms corresponding to these genes. Then we find the percentage z_i of common elements in S_i that are also present in S_j by

$$x_i = S_i \cap S_j \tag{2}$$

$$z_i = \frac{|x_i|}{|S_i|}. \tag{3}$$

Here z_i can assume values between zero and one. We would like to remark that the way we find the overlap is asymmetric, i.e., $z_i \neq z_i$ if $|S_i| \neq |S_j|$. That means the percentage overlap is taken with respect to the first signature set S_i.

From comparing the gene overlap (see Fig. 3A), the signature of Pei is the only one that is completely included in two other signatures namely Ben-porath-prc2 and Sotiriou-93. Interestingly, there is no unique signature, which means that each signature has some overlap with at least one other signature. The signature with the least commonality with other signatures is from Welm, which has only genes in common with the signatures of Taube and Reuter. Also the signature from Adorno has only a gene overlap with 4 other signatures. The signatures with the largest number of overlaps are Hua, Reuter and Sotiriou-93. These three signatures are sharing genes with 42 other signatures. This means that the overlap with other signatures varies considerably from 2 to 42. These numbers are added to Fig. 3A in the last column of the heat map.

In contrast to this, the overlap of GO-terms among signatures is shown in Fig. 3B. Also here the overlap among the signatures varies considerably. For instance, the signatures of Hua and Reuter share the highest overlap of 2614 GO-terms, whereas Adorno and He, Adorno and Welm have the lowest overlap of 1 GO-term. However, the most important result is that all signatures share at least some GO-terms with every other signature (see last column). Hence, all signatures have a non-zero overlap in their biological meaning as measured by GO-terms. This is different to the gene-overlap shown in Fig. 3A.

Hierarchy levels of GO-terms. For our last analysis of the signatures, we are mapping the GO-terms to structural features of a GO-DAG. Specifically, we obtain information about the hierarchy levels of the GO-terms.

In Fig. 3, we show the distributions for the hierarchy levels of the GO-terms of BP. This means that for each signature, the levels of the GO-terms of BP are obtained and a boxplot of the distribution is shown. Interestingly, a large number of signatures exhibit a similar distribution for the levels, and most of the signatures have the same median value of 7 (except for Wong-proteas, Tavazoie, Pei, Mori, Hu, Buffa, Ben-Porath-Prc2 and Adorno). Furthermore, all signatures, besides Wong-proteas, Welm, Tavazoie, Ivshina, Hu, and Glinsky, are symmetric.

Figure 2. (**A**) Overview of the total number of biomarker genes in each study. (**B**) Shown is the number of GO-terms in each study. The green points correspond to BP, the red points to MF and blue points to CC. (**C**) The percentage of GO-terms of BP, MF and CC used by each study. The color is the same as for B.

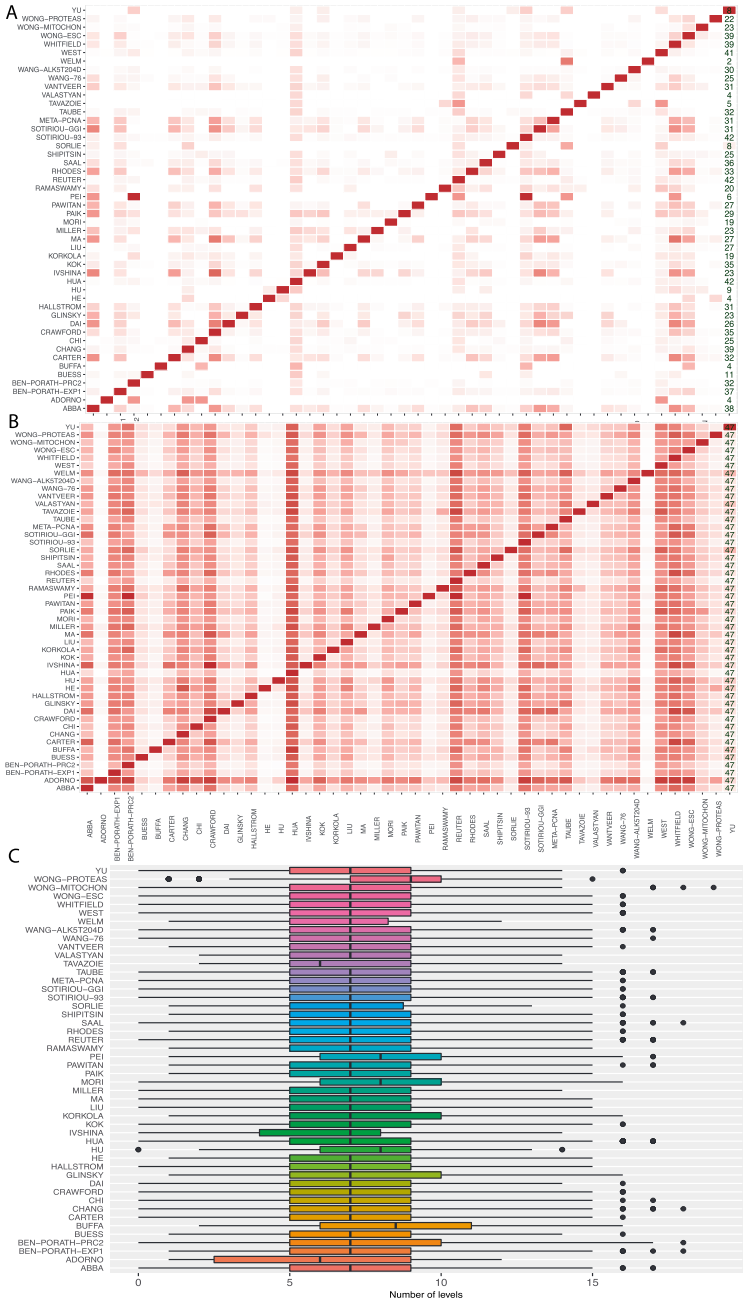


Figure 3. (**A**) Pairwise overlap of genes in BM signatures. (**B**) A: Pairwise overlap of GO-terms in BM signatures. (**C**) The distribution of GO-terms of BP hierarchy levels for each study.

Specifically, the signatures Wong-proteas, Welm, Ivshina and Hu are skewed to the right whereas the remaining ones are skewed to the left. The Wong-Proteas signature also has the highest median value (9), while Adorno and Tavazoie have the lowest median values (6). The degree of variation for the levels remains virtually the same for most of the signatures with the exception of a few.

These results demonstrate that despite the size differences of the signatures (see Fig. 2A), the differences in the number of GO-terms (see Fig. 2B) and the sparsity in the overlap of genes of the signatures (see Fig. 3A) the biological specificity of the GO-terms is very similar.

Prognostic prediction capabilities of random gene sets. In the following, we investigate the prognostic prediction capabilities of BM signatures and random gene sets systematically. We start by focusing on BM signatures and random gene sets for which the BM signatures have been removed. Thereafter, we investigate random gene sets for which not only the BM signatures have been removed but also further genes that share common biological processes. This will lead to more stringent insights about the biological meaning of BM signatures.

Effect of removing individual BM signatures. The study by[17] investigated prediction capabilities of random gene sets, RGS_i, whereas the genes in RGS_i were randomly sampled from the set $G_i' = G \setminus BM_i$. Here G corresponds to the total number of genes in our breast cancer data set and BM_i is the BM signature of study i, for $i \in \{1, \ldots, 48\}$. The number of genes sampled per random signature is the same as in BM_i, i.e., $|RGS_i| = |BM_i|$. We repeat this sampling 1000 times for each study, i.e., we studied 48, 000 random gene sets that have been constructed in this way.

We would like to remark, that the study by[17] did not apply a multiple testing correction to the obtained p-values despite the fact that multiple hypotheses had been tested. In order to see if these previous results are statistically robust, we repeated their analysis using a conservative Bonferroni correction[24]. Therefore, in total, we study 96, 000 random gene sets with and without Bonferroni correction.

The results of this analysis are shown in Fig. 4. Here the red/green points are the outcomes of the original BM signatures whereas dark red/dark green colors indicate non-significant values and light colors correspond to significant results. The violet distributions correspond to results from random signatures and the shaded green bars correspond to the lower 3*rd* percentile of these distributions. Furthermore, the horizontal black lines represent the median values of the distribution of random signatures and the long horizontal blue line corresponds to a significance level of $\alpha = 0.001$. Note that for the p-values a logarithmic scale (i.e., log_{10}) is used.

First, we observe from Fig. 4 that not all BM signatures (big points) lead to significant results. Specifically, the dark red and dark green points correspond to non-significant results whereas the light red and light green points correspond to significant results. This is a result from using different validation data than have been used by the original 48 BM studies. Still, without and with Bonferroni correction there are 39 BM signatures significant in each case. Hence, the remaining 9 signatures do not show prognostic value for independent validation data and lack robustness.

Furthermore, from Fig. 4 without a Bonferroni correction (left), we find that the median p-values of 37 studies are significant (11 studies are not significant) while for a Bonferroni correction (right), we find only 19 significant studies (29 studies are not significant). Also, we find that with and without a Bonferroni correction most lower 3% percentiles (green bars) are significant.

In order to obtain a better understanding of the total number of random gene signatures, we estimate an upper bound of the binomial coefficient $\binom{n}{k}$. Here n is the available number of genes and k is the size of a random biomarker set. The meaning of this binomial coefficient is the total number of random gene sets that can be formed by selecting k genes from all available n genes.

For our data set the order of magnitude of n is 10^4 and according to Fig. 2A the average size of a BM signature is $k = 10^2$. For the following estimate, only the order of magnitude of n and k are important as we will see below. Due to the fact that $\binom{10,000}{100}$ cannot be evaluated numerically, we estimate an upper bound of this by

$$\binom{n}{k} \leq \left(\frac{n \cdot e}{k}\right)^k. \tag{4}$$

Here e is Euler's number. The right-hand-side of Eq. 4 can be simplified by

$$\left(\frac{n \cdot e}{k}\right)^k = 10^x \Rightarrow \tag{5}$$

$$x = k \cdot \log_{10}\left(\frac{n \cdot e}{k}\right) \tag{6}$$

to obtain the order of magnitude as an exponent of 10. Overall, this leads to the following approximation of the binomial coefficient

$$\binom{n}{k} \leq \left(\frac{n \cdot e}{k}\right)^k = 10^x = 10^{243} \tag{7}$$

for values of $n = 10,000$ and $k = 100$ and x given in Eq. 6 (after rounding to integer numbers).

This demonstrates that the average number of random gene sets is in the order of 10^{243} and that one percentile of these correspond to 10^{241} different random gene sets, for each study. Hence, even for studies for which only

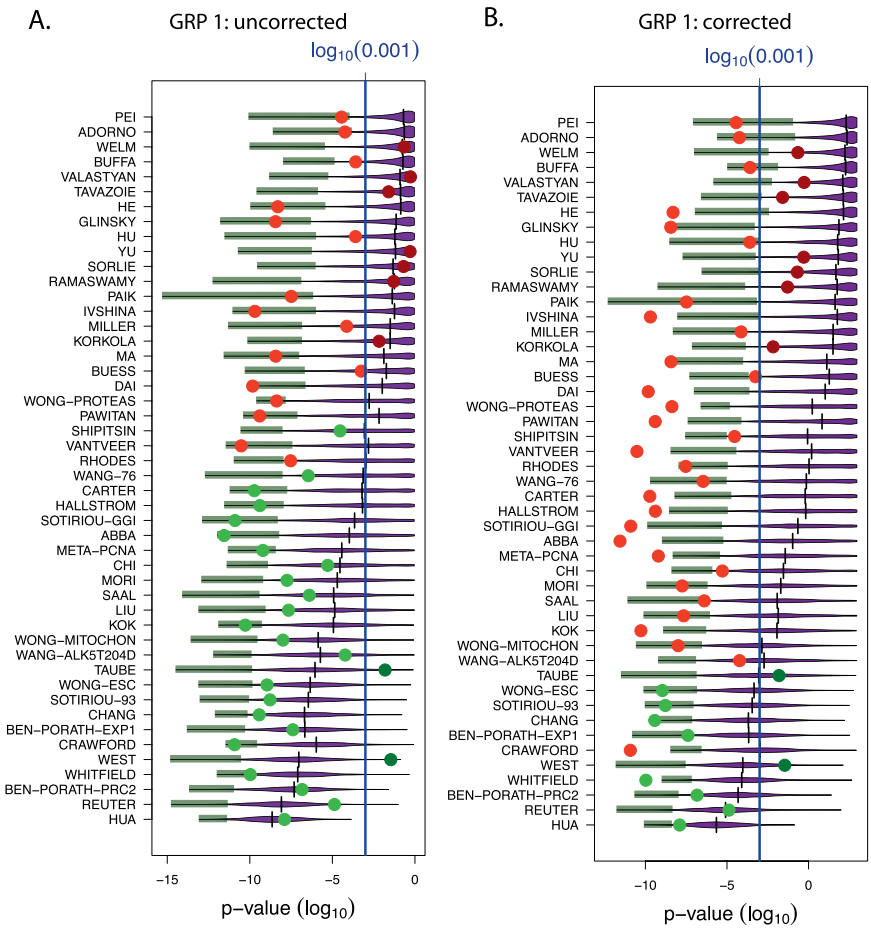


Figure 4. Results for gene removal procedure 1 for the NKI data. Shown are prognostic prediction capabilities of surrogate gene sets for 48 studies after removing BM signatures. Left: Results for uncorrected p-values (as in the original study[17]). Right: Bonferroni corrected p-values.

about three percent of all random gene sets are significant, corresponding to the lower 3*rd* percentile (green bars) in Fig. 4, the number of such gene sets is very large. In order to distinguish such significant random gene sets from non-significant gene sets we call the former *surrogate gene sets* because they have the same prognostic prediction capabilities as the BM signatures. Hence, the lower 3*rd* percentile corresponds to 10^{241} surrogate gene sets.

Effect of removing related biological processes. In our next analysis, we go one step further. Instead of only removing BM signatures, we remove also genes that belong to the same biological processes as the genes in the BM signatures (see GRP 2 in the Methods section). Due to the fact that according to the gene ontology (GO) database the biological processes are hierarchically organized, we approach this analysis iteratively by removing successively genes of biological processes on the same hierarchy level. Details of gene removal procedure 2 are described in the Methods section.

We assess the prediction results again by the p-values from the survival analysis. In addition, we assess the accuracy of predictions by declaring significant p-values as true positives (TPs) and non-significant results as false negatives (FNs). This allows the estimation of accuracy values. These evaluations are obtained for each hierarchy level.

In Table 1, we show three representative results for the signatures of Pei (top), Chang (middle) and Wong-Mitochon (bottom). The underlying p-values have been Bonferroni corrected. The remaining results for the remaining signatures can be found in the Tables 1 to 96 in the supplementary file. The first column shows the hierarchy level up to which the GO-terms have been removed (see GRP 2 Methods section) and columns two to six give further details about the involved genes and GO-terms. The accuracy (Acc) summarizes the results

Hierachy level	Genes removed	Cum. sum of genes removed	Genes left	GO-terms removed	Cum. sum of GO-terms removed	Acc. (%)	Sig.Acc. (%)
16	3	3	12,135	1	1	3.3	2.5
15	8	11	12,127	3	4	4.0	
14	2	13	12,125	1	5	2.9	
13	177	190	11,948	2	7	4.7	
12	1318	1508	10,630	9	16	3.3	
11	616	2124	10,014	8	24	3.7	
10	432	2556	9582	10	34	3.2	
9	229	2785	9353	11	45	3.4	
8	190	2975	9163	13	58	3.1	
7	197	3172	8966	9	67	2.8	
6	533	3705	8433	24	91	2.5	
5	854	4559	7579	16	107	3.2	
4	837	5396	6742	4	111	4.1	
3	130	5526	6612	4	115	2.5	
2	250	5776	6362	3	118	3.0	
1	56	5832	6306	1	119	3.2	
18	17	17	11,822	1	1	87.0	88.5
17	11	28	11,811	3	4	87.9	
16	3	31	11,808	2	6	89.6	
15	89	120	11,719	8	14	89.2	
14	203	323	11,516	24	38	89.8	
13	819	1142	10,697	41	79	87.4	
12	1932	3074	8765	91	170	87.1	
11	1246	4320	7519	112	282	84.8	
10	1216	5536	6303	153	435	84.3	
9	1326	6862	4977	174	609	81.9	
8	1252	8114	3725	183	792	79.6	
7	1258	9372	2467	193	985	70.8	
6	711	10,083	1756	165	1150	81.2	
5	573	10,656	1183	97	1247	73.2	
4	336	10,992	847	54	1301	88.1	
3	122	11,114	725	27	1328	86.3	
18	17	17	11,913	1	1	78.9	78.0
17	9	26	11,904	3	4	74.9	
14	63	89	11,841	9	13	76.7	
13	228	317	11,613	12	25	78.0	
12	1228	1545	10,385	25	50	78.7	
11	826	2371	9559	53	103	76.3	
10	1054	3425	8505	60	163	75.5	
9	1032	4457	7473	70	233	76.8	
8	949	5406	6524	77	310	78.6	
7	1754	7160	4770	89	399	69.7	
6	810	7970	3960	78	477	68.6	
5	868	8838	3092	67	544	63.4	
4	461	9299	2631	35	579	63.6	
3	339	9638	2292	24	603	65.6	
2	238	9876	2054	12	615	56.0	
1	30	9906	2024	3	618	56.6	

Table 1. Results for GRP 2 (NKI data) for three signatures. Top: Pei. Middle: Chang. Bottom: Wong–Mitochon. The p-values have been Bonferroni corrected.

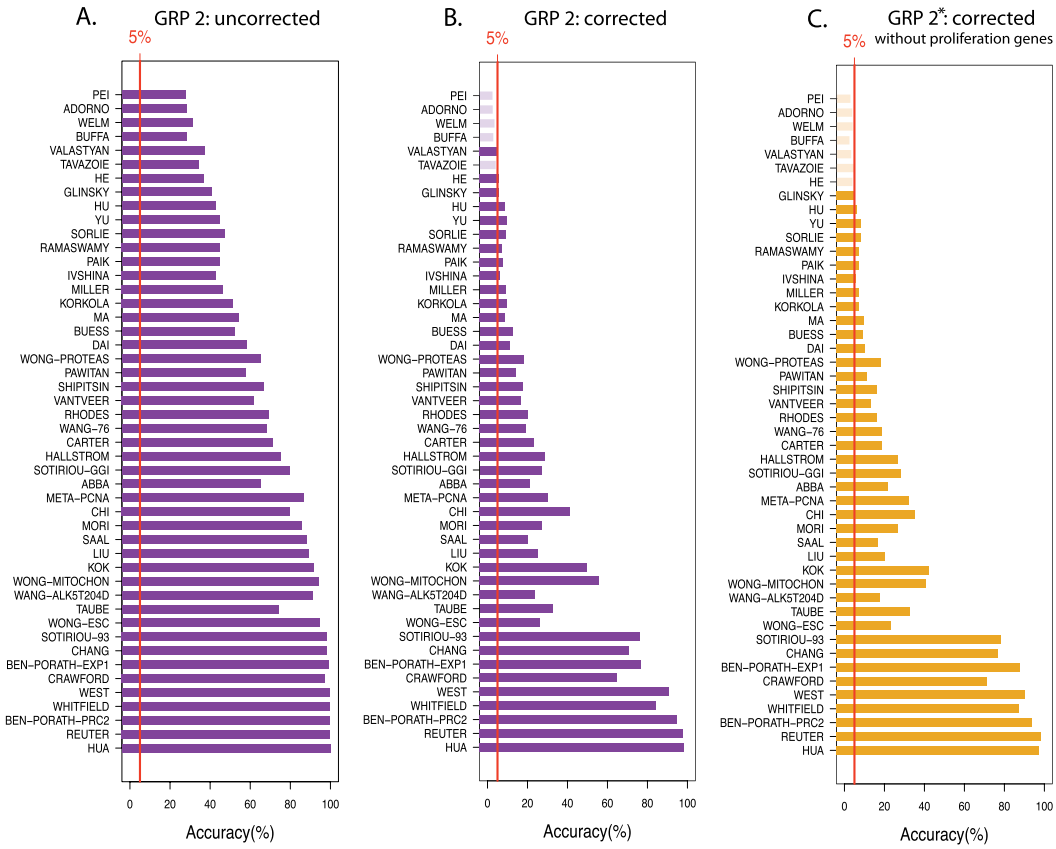


Figure 5. Results for gene removal procedure 2 for the NKI data. (**A**), (**B**) Show results for the minimal accuracy values across all hierarchy levels, whereas A is for uncorrected p-values and B for Bonferroni corrected p-values. (**C**) Results for GRP 2* for removing all GO-terms on all hierarchy levels (Bonferroni corrected).

of 1000 repeats and the overall significant p-values. Finally, Sig. Acc. gives the accuracy when only the signature is removed.

As one can see the accuracy values can be very low (Pei: top) or very high (Chang:middle) regardless of the hierarchy level, or they can decline toward higher hierarchy levels (Wong–Mitochon: bottom). However, despite this complicated behavior a commonality among all 48 studies is that there is always a non-vanishing percentage of random gene sets that make the correct predictions. Hence, the number of surrogate gene sets is non-zero for all signatures.

This is summarized in Fig. 5. Specifically, the shown accuracy values correspond to the minimal values for each study across all hierarchy levels. For instance, for Chang the minimal accuracy is 70.8% obtained for hierarchy level 7; see Table 1. From this figure one can see that also the resulting minimal accuracy values vary considerably across the studies, however, only 5 studies have values slightly smaller than 5%. All other studies have larger values than 5% and some are even larger than 80%, even for Bonferroni corrected p-values. Examples for the latter are the signatures from West, Whitfield, Ben-Porath-Prc2, Reuter and Hau.

For each hierarchy level of each study, one can investigate the resulting distribution of p-values for the random gene sets (similar to Fig. 4). Due to the fact that for each study many hierarchy levels have been studied (see Table 1 or the supplementary Tables 1 to 96) there are more than 1000 such distributions for all studies. For instance, for Pei there are 16 such distributions corresponding to 16 hierarchy levels (see Table 1). In order to simplify the presentation, we show only results for the minimal accuracy values in Fig. 5. The corresponding results are shown in Fig. 6. Interestingly, these results are qualitatively comparable to the results shown in Fig. 4. However, quantitatively, the difference is that *in average* these p-values are slightly larger. This implies, e.g., that the median values of less studies are significant. Specifically, in Fig. 6 the median values of 33 (without Bonferroni correction) respectively 10 (with Bonferroni correction) studies are significant.

In order to estimate the number of surrogate gene sets, we perform a similar approximation of the binomial coefficient as in Eq. 7, however, considering the reduced number of available genes. From the tables in the

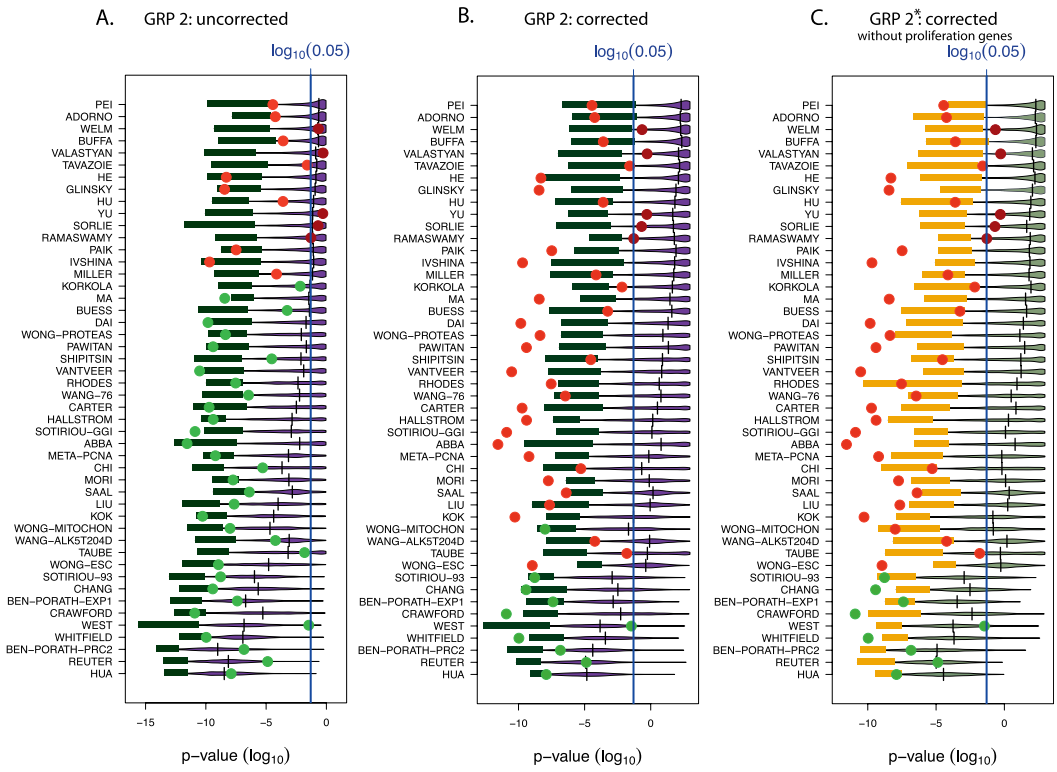


Figure 6. Results for gene removal procedure 2 for the NKI data. The results in (**A**) and (**B**) are for the minimal accuracy (see Fig. 5A,B), whereas A is for uncorrected p-values and B for Bonferroni corrected p-values. The results in (**C**) correspond to Fig. 5C where all GO-terms on all hierarchy levels have been removed and, in addition, all proliferation genes have been removed.

Supplementary File we observe, in average, $n = 1000$. Considering this, we obtain $x = 143$. Therefore, the total number of random gene sets constructed with GRP 2 is

$$\binom{n}{k} \leq 10^{143}. \tag{8}$$

Also this number is very large but a factor of 10^{100} smaller than the number of random gene sets obtained in Eq. 7.

From Fig. 5 and 6 (here the 3rd percentiles are highlighted in green) one can see that also for this procedure a certain percentile or random gene sets lead to the correct prediction outcome. Hence, the number of surrogate gene sets is for GRP 2 in the order of 10^{141}.

Finally, we repeat the above analysis for another data set from[18], in order to demonstrate the robustness of our results. In Figs. 7 and 8 we show results for the SWE data. Specifically, in Fig. 7 (top row) we use patient samples for LumA, LumB and Her2, in Fig. 7 (bottom row) LumA and Her2 and in Fig. 8 LumA and LumB. As one can see our results for the NKI data are confirmed for the SWE data for different subtypes of cancer. Other combinations of the subtypes give similar results (not shown).

Discussion

In this paper, we conducted a systematic study investigating the prognostic prediction capabilities of random gene sets. For this, we defined two different gene removal procedures (GRP 1 and GRP 2) for a constrained-sampling of random gene sets.

For clarity, we distinguish in our paper between three different types of genes set. The first one, called a signature, is a gene set identified in a targeted way. Typically, such genes are identified because it is assumed that they are biologically informative for a particular problem. In addition, if used in a prognostic prediction task such a signature yields statistically significant results, which evidences practically that the signature is indicative for the disease progression of patients. In contrast, a random gene set is obtained by randomly sampling genes

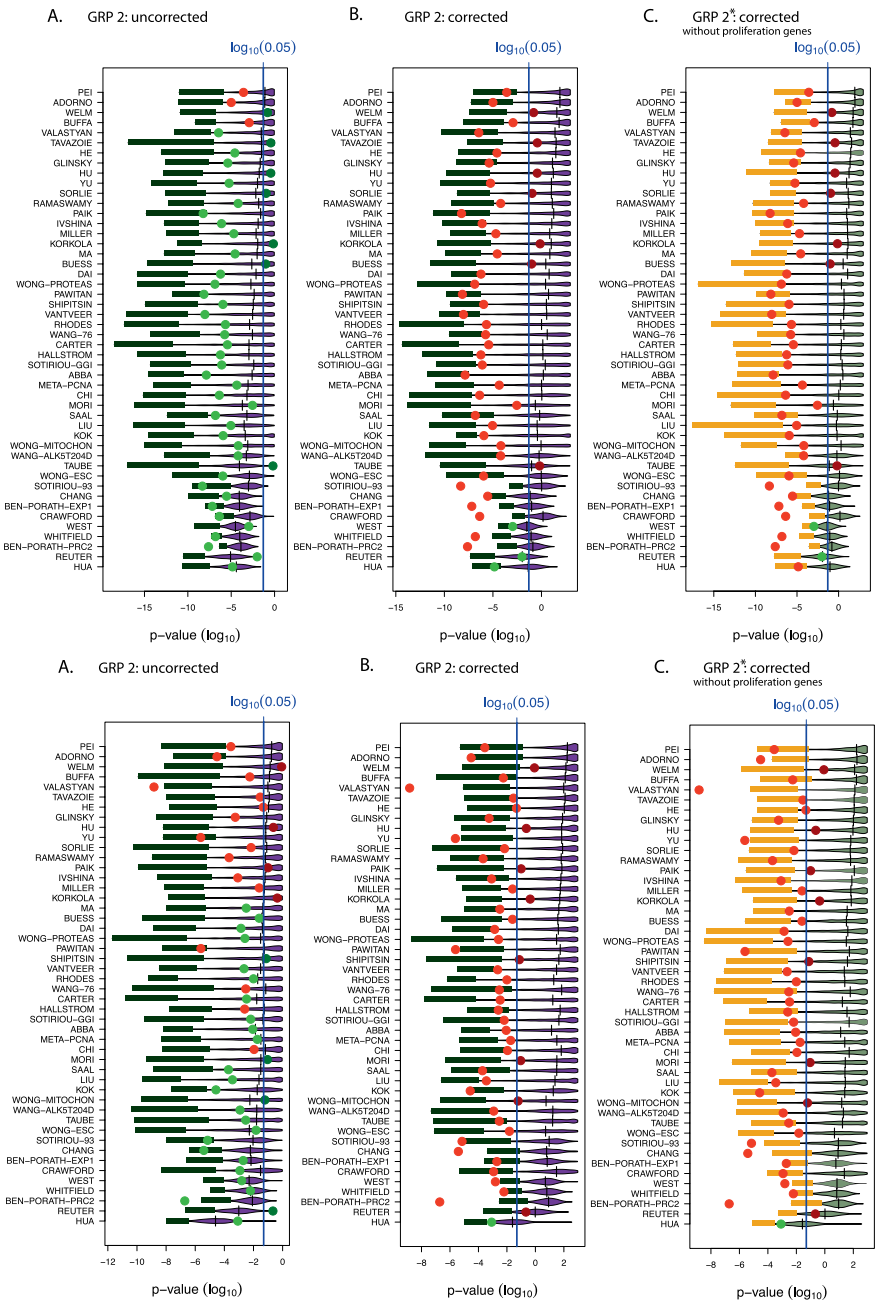


Figure 7. Results for gene removal procedure 2 and the SWE data similar to Fig. 6. Top row: Patient samples contain LumA, LumB, Basal and Her2. Bottom row: Patient samples contain LumA and Her2.

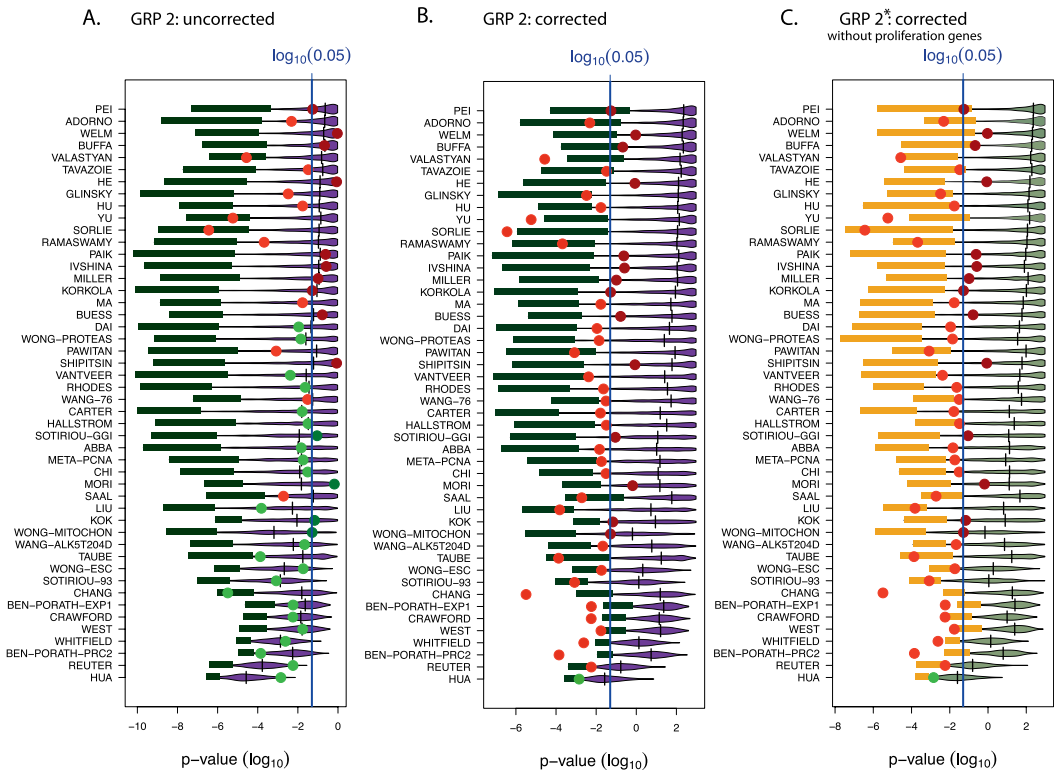


Figure 8. Results for gene removal procedure 2 and SWE data similar to Fig. 6. Patient samples contain LumA and LumB.

from an available gene pool. No particular meaning or role is attributed to such genes before sampling. Lastly, a surrogate gene set is a random gene set that has the same prognostic prediction capabilities as a signature. In our case this is indicated by a significant p-value from a survival analysis.

Results from gene removal procedure 1. The results from GRP 1 are summarized in Fig. 4. From this figure, one obtains the following interpretations.

Most random signatures are significantly associated with prognostic outcome: This is only correct for random signatures with a median value that is statistically significant, because the median corresponds to 50% of the population. Hence, a significant median indicates that 50% of the surrogate signatures are significant. In summary, without a Bonferroni correction this is correct for 37 (77.1%) studies and with Bonferroni correction for 19 (39.6%) studies (see Fig. 4). Hence, this statement is signature-dependent and does not hold generally.

Many surrogate signatures are significantly associated with prognostic outcome: This statement is correct for surrogate signatures for which a certain percentile of the surrogate signatures is statistically significant. From Fig. 4 one can see that this is correct for the lower 3rd percentiles for all studies, with and without a Bonferroni correction.

The number of surrogate signatures which are significantly associated with prognostic outcome is very large: Despite the fact that not all median values for all signatures are significant, the number of surrogate signatures that are significant is for each study very lage. This result has been obtained from approximating the upper bound of the binomial coefficient $\binom{n}{k}$ where n is the number of available genes and k is the size of a surrogate gene set. As an approximation we found

$$\binom{n}{k} \leq 10^{243} \tag{9}$$

for values of $n = 10,000$ and $k = 100$.

Results from gene removal procedure 2. The results from GRP 2 are an extension of GRP 1 in the sense that the genes available for random sampling are further constricted. That means instead of only removing BM

signatures, in addition, also genes related to the same biological processes are removed. Hence, the random gene sets obtained from this procedure are less biologically similar to the BM signatures.

The initial motivation for exploring GRP 2 came from the observation that the overlap of biological processes present in random gene sets and in BM signatures is non-zero. That means whenever at least one gene in a random gene set belongs to a GO-term that is also present for a BM signature, possibly for a different gene, the random gene set and the BM signature have this GO-term in common. Numerically, we find the average number of common GO-terms (corresponding to BP) across all signatures is 341. We find the largest overlap for Hua with 2602 and the smallest for Adorno with 1 GO-term. These numbers are understandable because Hua contains the largest number of GO-terms (over 5000) whereas Adorno contains the smallest number (19 GO-terms of BP); see Fig. 2.

The results from GRP 2 for the NKI data are summarized in Figs. 5 and 6 (and for the SWE data in Figs. 7 and 8). It is interesting to note that the qualitative results are similar for GRP 2 and GRP 1. That means even by removing genes related to the same biological processes as the signature genes, the prognostic prediction capabilities of surrogate gene sets can be confirmed. Importantly, qualitatively, GRP 1 and GRP 2 are entirely different with respect to their biological meaning. Specifically, we designed GRP 2 in a way that the procedure allows the gradual removal of more and more *biological meaning* from random gene sets. This is accomplished by a ranking of GO-terms according to their hierarchy levels because it is know that GO-terms in a GO-DAG on higher levels contain biological information that is more specific than GO-terms on lower levels[25]. Due to the fact that GRP 2 removes genes, associated with certain GO-terms, gradually from high to low hierarchy levels, we were able to study this effect explicitly; see Table 1 and Tables 1 to 96 in the Supplementary File (for the NKI data). We would like to remark that removal of GO-terms from all hierarchy levels (corresponding to the last step of GRP 2) results in random gene sets with no biological similarity to the original BM signature. Hence, per construction, such random gene sets have a biological similarity of zero with the original BM signature.

Considering the biological differences in the meaning of random genes sets resulting from GRP 1 and GRP 2 the results in Figs. 5 and 6 are remarkable because it means that any biological justification given for the selection of an original BM signature is anecdotal. Specifically, by removing genes related to the same biological processes as the signature genes we eliminate the possibility of *accidentally* selecting genes for a random gene set that share the same biological interpretation as the original BM signature. Hence, any biological interpretation of such a BM signature is meaningless because we demonstrated that one can find surrogate gene sets with the same prediction capability but entirely different biological interpretations due to zero overlap in the GO-terms of involved genes. This is also true for GRP 2* where additionally proliferation genes have been removed (see Fig. 6C).

We would like to remark that the study by[17] did not allow this conclusion because BM signatures have not been removed nor genes from associated biological processes. This leaves the possibility of *accidentally* selecting genes for a random gene set that share the same biological interpretation as the original BM signatures because these genes belong to the same biological processes as indicated by common GO-terms in the domain BP.

Our study is also different to[21] where the investigation by[17] has been extended by removing proliferation genes. The problem with their design is that resulting random gene sets can still have a non-vanishing overlap of common GO-terms and, hence, share to a certain extend biological meaning with a signature. Instead, we aimed at the elimination of all common GO-terms so that the resulting random gene sets have a different biological meaning. Further constraining of GRP 2 by additionally removing proliferation genes, as studied in[21], which we named GRP 2*, does not change our main result.

Taking a more specific look into some of the studies we used for our analysis allows to make this point more clear. For instance, the study by[26] identified a BM signature by computationally investigating 42 breast cancer gene expression studies. After demonstrating the prognostic capability of their signature the biological importance of these genes has been discussed and their functional role has been characterized as cell cycle process related and response to steroid hormone stimulus. Similarly, in the studies by Carter[27], Chi[28], Saal[29], Shipitsin[30] and West[31] the biological importance of their signatures pointed to chromosomal instability, hypoxia response, PI3K pathway signaling, TGF-β signaling pathway and stromal response respectively. However, based on our results, none of these biological interpretations established a causal explanation of the underlying cancer biology because one can always find alternative gene sets, which we called surrogate gene sets, that contain neither genes from their signatures nor from genes with related biological processes (nor from proliferation genes) but achieve the same prognostic predictions.

From these and other studies, one can derive the following general pattern that can be found in many prognostic breast cancer studies. First, signature genes are identified by computational, experimental or mixed-approaches and, second, the biological relevance of the signature genes is discussed. Our results demonstrate that neither step is necessary. The first step can be omitted because we showed that a constrained random sampling can lead to surrogate gene sets with the same prognostic prediction capabilities. Hence, any sophisticated, e.g., biology-driven selection process is equivalent to a random selection process. From our analysis we found that the probability that such a random gene set is actually a surrogate gene set is in the percentage range.

The second step can be omitted because we showed that by GRP 2 one can systematically construct surrogate gene sets with an entirely different biological meaning as the signature genes. Specifically, due to the fact that we remove systematically all genes related to any biological process of the signature genes, none of the genes in a surrogate gene set can belong to any of these biological processes. Formally, this can be written as follows (for the removal of all hierarchy levels). For any signature

$$BM = \{g_1, \ldots, g_m\} \tag{10}$$

consisting of m genes and corresponding GO-term set

$$GT = \{GO_1, \ldots, GO_t\} \tag{11}$$

representing all GO-terms of the genes in *BM* and any surrogate gene set

$$SGS = \{g'_1, \ldots, g'_m\} \tag{12}$$

consisting of m genes and corresponding GO-term set

$$GT' = \{GO'_1, \ldots, GO'_{t'}\} \tag{13}$$

representing all GO-terms of the genes in *SGS*, with t possibly different to t', the two sets GT and GT' are disjoint, i.e.,

$$GT \cap GT' = \emptyset. \tag{14}$$

For our results shown in the Tables 1 to 144 (Supplementary File), including also the removal of proliferation genes, this holds for the last row in these tables, i.e., the highest level. Hence, due to the fact that the signature genes (i.e., *BM*) and the surrogate gene set (i.e., *SGS*) do not share any GO-term they have a complementary biological meaning. Furthermore, there is not just one surrogate gene set but in the order of 10^{141} different sets. This demonstrates that the biological discussion of *BM* is meaningless because one can find a huge number of surrogate gene sets with a plurality of biological meanings.

In contrast to studies investigating the problem of reproducibility of biomedical results[32] requiring the adjustment of approaches, our paper is different because our results point to a fundamental lack of a commonly used framework which is unfixable. As a generalization of our results for 48 signatures, we assert that a signature with a sensible biological interpretation cannot be found within the studied prognostic framework utilizing survival analysis. More formally, this means the commonly used prognostic framework is no causal model[33].

In conclusion, we demonstrated that the common assumption that "A reliable set of predictive genes also will contribute to a better understanding of the biological mechanism of metastasis"[7] is not true.

Falsification mechanism to test biological meaning of prognostic signatures. For testing the validity of general signatures, we suggest the following procedure to test if it is justified to investigate the biological meaning of a prognostic signature of breast cancer.

1. G : total number of genes in a breast cancer dataset.

2*. *Optional step: Removing proliferation genes in *PG* from G. The set *PG* contains proliferation genes. This gives a new set of genes G^* with $G^* = G \setminus PG$.

2. $BM : \{g_1, \ldots, g_m\}$. *BM* is the gene signature and $g_1, \ldots, g_m$ are the genes in the corresponding signature.

3. Mapping of the genes in *BM* to GO-terms. This gives:

$$BM = \{g_1, \ldots, g_m\} \rightarrow \{GO_1, \ldots, GO_t\}. \tag{15}$$

Note, each gene can be connected to more than one GO-term. For this reason $m \leq t$.

4. Mapping of the GO-terms to genes. This gives:

$$GO_i \rightarrow g(i) = \{g_1(i), \ldots, g_k(i)\}. \tag{16}$$

for all GO-terms i with $i \in \{1, \ldots, t\}$.

5. Delete all the genes in $D = \cup_{i \in \{1,\ldots,t\}} g(i)$ from G. This results in a new gene set given by $G' = G \setminus D$.

6. From G', sample new sets of random genes of size $|BM|$ and perform the prognostic task. This is repeated 1000 times.

7. Application of a Bonferroni correction to the p-values and assessing the performance for a significance level of α.

From numerical analyses, we found that 1000 repeats are sufficient to estimate the tail distribution of random gene sets because, for the signatures studied in this paper, the *probability to be a surrogate gene set* (p_{sgs}) is in average 3% percent or higher. However, other signatures may require larger repeats due to the reciprocal relation between these entities, i.e., # repeats $> 1/p_{sgs}$.

If this procedure does not result in any surrogate gene set with the same prognostic prediction capabilities, the BM signature has a biological meaning that deserves to be discussed. Otherwise the BM signature has no sensible biological interpretation, which is the case for the 48 signatures studied in this paper.

Conclusion

In this paper, we shed light on the biological interpretability of BM signatures for the prognostic prediction of breast cancer. Our results demonstrate that none of the 48 studied signatures has a sensible biological interpretation because for each, surrogate gene sets can be found that perform the same task, however, belonging to different biological processes. This implies that every signature (random or not) can just serve as a *black-box* prediction model without a biological interpretation. We believe that this has wider implications, even beyond biomedicine, to general machine learning and artificial intelligence models but this remains to be studied[34]. In addition, we proposed a procedure to test the biological meaning of prognostic signatures of breast cancer. This test could avoid further confusion in the literature about the biological meaning of prognostic signatures.

It is widely know that prognostic signatures of breast cancer are very heterogeneous and sensitive to changes in the studied perspective. For this reason, we assumed in this paper a higher conceptual ground, based on a systems-view, in order to study a common aspect shared by many signatures that allows to pierce through the unavoidable variability and heterogeneity. This concept goes back to the roots of systems biology as envisioned in[35,36].

Received: 16 July 2020; Accepted: 3 December 2020
Published online: 08 January 2021

References

1. Idris, S. F., Ahmad, S. S., Scott, M. A., Vassiliou, G. S. & Hadfield, J. The role of high-throughput technologies in clinical cancer genomics. *Exp. Rev. Mol. Diagn.* **13**, 167–181 (2013).
2. Cohrs, R. J. *et al.* Translational medicine definition by the european society for translational medicine. *New Horizons Transl. Med.* **2**, 86–88. https://doi.org/10.1016/j.nhtm.2014.12.002 (2015).
3. Bullinger, L. *et al.* Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N. Engl. J. Med.* **350**, 1605–1616 (2004).
4. Simon, R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br. J. Cancer* **89**, 1599–1604 (2003).
5. Kim, C. & Paik, S. Gene-expression-based prognostic assays for breast cancer. *Nat. Rev. Clin. Oncol.* **7**, 340 (2010).
6. Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: A multiple random validation strategy. *The Lancet* **365**, 488–492 (2005).
7. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci.* **103**, 5923–5928 (2006).
8. Haury, A.-C., Gestraud, P. & Vert, J.-P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE* **6**, 1 (2011).
9. Gilhodes, J. *et al.* Comparison of variable selection methods for high-dimensional survival data with competing events. *Comput. Biol. Med.* **91**, 159–167 (2017).
10. Kim, S.-Y. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinform.* **10**, 147 (2009).
11. Van De Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
12. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* **365**, 671–679 (2005).
13. Liu, H. *et al.* High-dimensional semiparametric gaussian copula graphical models. *Ann. Stat.* **40**, 2293–2326 (2012).
14. Domany, E. Using high-throughput transcriptomic data for prognosis: A critical overview and perspectives. *Cancer Res.* **74**, 4612–4621 (2014).
15. Vidal, M. A unifying view of 21st century systems biology. *FEBS Lett.* **583**, 3891–3894 (2009).
16. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
17. Venet, D. *et al.* Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **7**, e1002240 (2011).
18. Brueffer, C. *et al.* Clinical value of rna sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter sweden cancerome analysis network?breast initiative. *JCO Precis. Oncol.* **2**, 1–18 (2018).
19. Emmert-Streib, F. & Dehmer, M. Introduction to survival analysis in practice. *Mach. Learn. Knowl. Extract.* **1**, 1013–1038. https://doi.org/10.3390/make1030058 (2019).
20. Manjang, K., Tripathi, S., Yli-Harja, O., Dehmer, M. & Emmert-Streib, F. Graph-based exploitation of gene ontology using GOxploreR for scrutinizing biological significance. *Sci. Rep.* **10**, 1–16 (2020).
21. Goh, W. W. B. & Wong, L. Why breast cancer signatures are no better than random signatures explained. *Drug Discov. Today* **23**, 1818–1823 (2018).
22. Whitfield, M. L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).
23. Emmert-Streib, F., Moutari, S. & Dehmer, M. A comprehensive survey of error measures for evaluating binary decision making in data science. *Wiley Interdiscipl. Rev.* e1303 (2019).
24. Emmert-Streib, F. & Dehmer, M. Large-scale simultaneous inference with hypothesis testing: Multiple testing procedures in practice. *Mach. Learn. Knowl. Extract.* **1**, 653–683. https://doi.org/10.3390/make1020039 (2019).
25. Dennis, G. *et al.* DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, R60 (2003).
26. Abba, M. C., Lacunza, E., Butti, M. & Aldaz, C. M. Breast cancer biomarker discovery in the functional genomic age: A systematic review of 42 gene expression signatures. Biomarker insights **5**, BMI–S5740 (2010).
27. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**, 1043–1048 (2006).
28. Chi, J.-T. *et al.* Gene expression programs in response to hypoxia: Cell type specificity and prognostic significance in human cancers. *PLoS Med.* **3**, 1 (2006).
29. Saal, L. H. *et al.* Poor prognosis in carcinoma is associated with a gene expression signature of aberrant pten tumor suppressor pathway activity. *Proc. Natl. Acad. Sci.* **104**, 7564–7569 (2007).
30. Shipitsin, M. *et al.* Molecular definition of breast tumor heterogeneity. *Cancer Cell* **11**, 259–273 (2007).
31. West, R. B. *et al.* Determination of stromal signatures in breast carcinoma. *PLoS Biol.* **3**, e187 (2005).
32. Begley, C. G. & Ioannidis, J. P. Reproducibility in science: improving the standard for basic and preclinical research. *Circul. Res.* **116**, 116–126 (2015).
33. Pearl, J. *Causality: Models, Reasoning, and Inference* (Springer, Cambridge, 2000).
34. Emmert-Streib, F., Yli-Harja, O. & Dehmer, M. Explainable artificial intelligence and machine learning: A reality rooted perspective. *WIREs Data Mining Knowl. Discov.* **10**, e1368. https://doi.org/10.1002/widm.1368 (2020).
35. Waddington, C. *The Strategy of the Genes* (Geo, Allen & Unwin, London, 1957).
36. Kauffman, S. *Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, Oxford, 1993).

Acknowledgements

Kalifa Manjang is supported by Tampere University via the Prostate Cancer Center. Matthias Dehmer thanks the Austrian Science Funds for supporting this work (Project P30031).

Author contributions

F.E.S. conceived the study. K.M., S.T. and F.E.S. conducted the analysis. K.M., S.T., O.Y.H., M.D., G.G. and F.E.S. interpreted the results. All authors wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-79375-y.

Correspondence and requests for materials should be addressed to F.E.-S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PUBLICATION
III

Limitations of explainability of Prognostic biomarkers of prostate cancer

K. Manjang, O. Yli-Harja, M. Dehmer and F. Emmert-Streib

Frontiers in Genetics 12.(2021), 1095
DOI: 10.3389/fgene.2021.649429

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 July 2021
doi: 10.3389/fgene.2021.649429

Limitations of Explainability for Established Prognostic Biomarkers of Prostate Cancer

Kalifa Manjang[1], *Olli Yli-Harja*[2,3,4], *Matthias Dehmer*[5,6,7] *and Frank Emmert-Streib*[1,4*]

[1] *Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland,* [2] *Computational Systems Biology, Tampere University, Tampere, Finland,* [3] *Institute for Systems Biology, Seattle, WA, United States,* [4] *Faculty of Medicine and Health Technology, Institute of Biosciences and Medical Technology, Tampere University, Tampere, Finland,* [5] *Department of Computer Science, Swiss Distance University of Applied Sciences, Brig, Switzerland,* [6] *Department of Mechatronics and Biomedical Computer Science, University for Health Sciences, Medical Informatics and Technology (UMIT), Hall, Austria,* [7] *College of Artificial Intelligence, Nankai University, Tianjin, China*

High-throughput technologies do not only provide novel means for basic biological research but also for clinical applications in hospitals. For instance, the usage of gene expression profiles as prognostic biomarkers for predicting cancer progression has found widespread interest. Aside from predicting the progression of patients, it is generally believed that such prognostic biomarkers also provide valuable information about disease mechanisms and the underlying molecular processes that are causal for a disorder. However, the latter assumption has been challenged. In this paper, we study this problem for prostate cancer. Specifically, we investigate a large number of previously published prognostic signatures of prostate cancer based on gene expression profiles and show that none of these can provide unique information about the underlying disease etiology of prostate cancer. Hence, our analysis reveals that none of the studied signatures has a sensible biological meaning. Overall, this shows that all studied prognostic signatures are merely black-box models allowing sensible predictions of prostate cancer outcome but are not capable of providing causal explanations to enhance the understanding of prostate cancer.

Keywords: prostate cancer, biomarkers, prognostic biomarkers, survival analysis, data science, computational biology, biostatistics

OPEN ACCESS

Edited by:
Natalia Polouliakh, Sony Computer Science Laboratories, Japan

Reviewed by:
Tianshou Zhou, Sun Yat-sen University, China
Padhmanand Sudhakar, KU Leuven, Belgium

****Correspondence:***
Frank Emmert-Streib
v@bio-complexity.com

Specialty section:
This article was submitted to Systems Biology Archive, a section of the journal Frontiers in Genetics

Received: *04 January 2021*
Accepted: *01 June 2021*
Published: *22 July 2021*

Citation:
Manjang K, Yli-Harja O, Dehmer M and Emmert-Streib F (2021) Limitations of Explainability for Established Prognostic Biomarkers of Prostate Cancer. Front. Genet. 12:649429. doi: 10.3389/fgene.2021.649429

1. INTRODUCTION

Prostate cancer (PCa) is the second most prevalent cancer among men, the average age of diagnosis is 66 years, and about 60% of diagnosed cases occur in men over 65 years old. In the United States, for example, 191, 930 newly diagnosis cases of PCa are estimated in 2020, resulting in about 33, 330 mortalities (Siegel et al., 2020). A substantial proportion of PCa is characterized as slow-growing and indolent requiring no immediate therapeutic intervention. However, tumor stages T1 and T2, and tumor stages higher than T2 are more aggressive and invade the surrounding organs and the patient is more likely to die from the disease (Chen et al., 2020). Specifically, for men with local or regional PCa, the 5-year survival rate is almost 100%, whereas the 5-year survival rate for men with metastatic PCa is 31%.

Since the inception of high-throughput technologies, a large number of molecular markers have been described in the literature capable of distinguishing cancer patients with good and bad prognosis. Nonetheless, few found their way into clinical decision making. Many biomarker studies have used genome-wide gene expression analysis to define unique gene expression signatures related to the prognosis of PCa. For example, Chen et al. (2012) developed a 7-gene prognostic signature through a cluster-correlation analysis to identify differentially expressed genes in various cell types associated with PCa progression. Likewise, in Liu et al. (2007), the gene expression of CD44+CD24 of low tumorigenic breast and normal breast epithelium cells were compared. They used the differentially expressed genes to construct a 186-gene "invasiveness" gene signature. The signatures were tested for their association with two clinical endpoints, overall survival and metastasis-free survival, in breast and other cancer patients. Interestingly, the signature was substantially correlated with the two survival endpoints in patients with breast cancer and other types of cancer. Another study by Ramaswamy et al. (2003) examined the molecular variations between human primary tumors and metastases. The gene expression profiles of different types of adenocarcinoma metastases and unmatched primary adenocarcinomas were compared, and the analysis identified a gene expression signature capable of separating primary from metastatic adenocarcinomas (Ramaswamy et al., 2003).

There are also studies that use more advanced approaches to derive the gene signatures. In a study by Irshad et al. (2013), a 19-gene signature enriched in indolent prostate tumors was identified. Their final signature includes three genes that, through a further classification of the 19-gene signature, was established by a decision tree (DT) model. Similarly, a combination of artificial neural network analysis and data from literature search and other studies resulted in a panel of PCa progression markers, which were used in a transcriptomic analysis of 29 radical prostatectomy samples correlated with clinical outcome (Larkin et al., 2012).

Aside from such potential success stories, there are several well-known problems with prognostic signatures. One such problem relates to the stability of the selection of prognostic genes. In Michiels et al. (2005), this has been studied for various cancer types and the authors found that the size of the training data as well as the patient data both crucially effect the selection of such genes. For breast cancer, this effect has been quantified by Ein-Dor et al. (2006). Specifically, the authors showed that thousands of patient samples are needed for achieving an overlap of 50% between two predictive sets of prognostic genes. Further examples of such studies reporting similar results can be found in Kim (2009), Haury et al. (2011), and Gilhodes et al. (2017). A well-recognized study by Venet et al. (2011) addressed yet another problem by showing that many random breast cancer gene sets have similar prognostic prediction capabilities as biomarker (BM) signatures. The study by Goh and Wong (2018) extended this by removing proliferation genes. A conceptual problem of both studies is that random gene sets could still share biological similarity on the level of biological processes (BPs). The reason for this is that no systematic mechanism has been implemented that would eliminate such a similarity. In contrast, the study by Manjang et al. (2021) introduced a gene removal procedure (GRP) that accomplished this.

The purpose of this paper is to test a hypothesis about the systems behavior of PCa. Specifically, despite well-documented differences between breast cancer and PCa, e.g., PCa affects men exclusively, whereas breast cancer commonly affects women, likewise both tumors arise in different organs involving different physiological functions, we hypothesize that their functional similarity, e.g., via the hallmarks of cancer (Hanahan and Weinberg, 2000, 2011), induces similar results for prognostic signatures. In order to investigate this, we study 32 published prognostic PCa signatures from the literature and demonstrate that random gene sets can be found with similar prediction capabilities as these signatures but opposite biological meaning.

The paper is organized as follows. In the next section, we describe our methods and data used for our analysis. Then we present and discuss our results. The paper completes with concluding remarks.

2. MATERIALS AND METHODS

In this section, we provide information about the data and methods used for our analysis.

2.1. Biomarker Signatures

We identified reported PCa gene signatures from a literature search. From this search, we found 32 signatures from 31 studies that have been published between 2002 and 2020. For all signatures, the Entrez gene IDs corresponding to the HGNC gene symbols are determined. All genes without an associated Entrez gene ID are discarded. **Table 1** shows an overview of the published gene signatures we use for our study.

2.2. Gene Expression Data

We collected RNA-seq data (HTSeq-FPKM and HTSeq-FPKM-UQ) of patients with PCa from the TCGA-PRAD project. We obtained the data from the UCSC Xena GDC data hub (https://xenabrowser.net/datapages/) on September 7, 2020. FPKM stands for *Fragments Per Kilobase of transcript per Million* mapped reads (Trapnell et al., 2010). It accounts for a situation in which only 1 end of a pair-end read is mapped. The FPKM of a gene is estimated as follows:

$$FPKM = \frac{10^9 \times \text{number of reads mapped to the gene}}{(\text{number of reads mapped to all protein}-\text{coding genes} \times \text{length of the gene in base pairs})} \quad (1)$$

Similarly, FPKM-UQ means Fragments Per Kilobase of transcript per Million mapped reads upper quartile. It is a modified estimate of FPKM where the total protein-coding read count is replaced by the 75th percentile read count for a sample. A notable difference between the two is the values of FPKM-UQ tends to be much higher due to the significant disparity between the total mapped number of reads in an alignment and the mapped number of reads to one gene.

The gene expression data set used in our study contains 551 samples, of which 498 are primary solid tumors, 52 are solid tissue normal, and one is metastatic. We exclude the metastatic and solid tissue normal samples from the data set. From these data, we used only protein-coding genes without missing information for the HTSeq-FPKM data cohort. Likewise, from the HTSeq-FPKM-UQ data we used only genes with $< 2\%$ missing information across all samples. The final HTSeq-FPKM data set contains 498 samples and 16,428 genes, whereas the HTSeq-FPKM-UQ data set contains 498 samples and 15,165 genes. Lastly, patient survival information for each sample was derived from Liu et al. (2018). Specifically, we used the progression-free interval end-points. In this paper, we refer to the HTSeq-FPKM and HTSeq-FPKM-UQ in our analysis as GDC cohort A and GDC cohort B, respectively.

2.3. Outcome Association

In order to determine the prognostic importance of a random gene set, we perform a survival analysis. We estimate Kaplan–Meier survival curves and compare these with a Mantel–Haenszel test (Emmert-Streib and Dehmer, 2019). That means each comparison provides a p-value from such a hypothesis test.

The patients are stratified into two classes (low and high risk) by using the PC1 method. This method categorizes patients according to a particular gene set. Hence, the resulting survival analysis is a function of the gene set used to categorize the patients. Overall, our study consists of three main steps: first, the selection/construction of random gene set; second, the classification of patients samples; and third, the survival analysis.

In the next section, we explain our method we use as GRP for constructing random gene sets.

2.4. Gene Removal Procedure

Our GRP entails the removal of both the BM signatures and genes that belong to the same BPs as the genes in the BM signatures. The gene ontology (GO) is hierarchical (Ashburner et al., 2000). Hence, we approach this analysis iteratively by removing genes of BPs successively from the same hierarchy level. The GRP we use is defined as follows:

1. G: the genes in the PCa data set (16,425 and 15,165 for GDC cohort A and B, respectively).
2. $BM_i : g_i, \ldots, g_m$. BM_i is the gene signature i (i range from 1 to 32) and $g_i, \ldots, g_m$ are the genes in the respective signatures.
3. Removing biomarker genes in signature BM_i from G. This produces a new set of genes G_i' with $G_i' = G \backslash BM_i$.

3* Optional step: Remove the proliferation genes, PG from G. This gives a new set of genes $G_i'^*$ with $G_i'^* = G' \backslash PG$.

4. Map the genes in BM_i to GO-terms and the corresponding hierarchy levels. This gives: $BM_i = \{g_i, \ldots, g_m\} \rightarrow R = \{(GO_1, L_1), \ldots, (GO_t, L_t)\}$(Manjang et al., 2020).

   Note, each gene can be connected to more than one GO-term. For this reason, $m \leq t$.
5. Map each GO-term in R, i.e., GO_i with $i \in \{1, \ldots, t\}$, to its gene set GS_i.
6. For each biomarker set i: Loop-over the elements in set R.
   a. Delete all the genes associated with the GO-terms in set R. This results in a new set given by $G'' = G' \backslash D$, where D is the set of genes having GO-terms in R, i.e., $D = \cup_{i=1}^{t} GS_i$.
7. From G'', we loop from 1 to 1,000:
   a. We sample new sets of biomarker genes of size $|BM_i \in G|$ and perform the prognostic task. We repeat this for 1,000 times.
   b. Application of a Bonferroni correction to the p-values.
   c. Set $G' = G''$. Stop if $l = L_{min}(i)$ or $|G''| < 2 \times |BM_i \in G|$.

In the above procedure, the optional step called 3* involves the removal of the 664 genes that are related to proliferation (this gene set is called PG). We extracted the genes in PG from Goh and Wong (2018).

TABLE 1 | Overview of published and evaluated prognostic signatures for prostate cancer used for our study.

Acronym of a study	Number of genes*	Cancer type	Reference
AGELL	12	Prostate cancer	Agell et al., 2012
BIBIKOVA	16	Prostate cancer	Bibikova et al., 2007
BISMAR	12	Prostate cancer	Bismar et al., 2006
CHEN	4	Prostate cancer	Chen et al., 2020
CHEN_CC	7	Prostate cancer	Chen et al., 2012
CHEVILLE	2	Prostate cancer	Cheville et al., 2008
CHU	8	Prostate cancer	Chu et al., 2018
CUZICK	31	Prostate cancer	Cuzick et al., 2011
GLINSKY	11	Multiple cancers	Glinsky et al., 2005
IRSHAD	19	Prostate cancer	Irshad et al., 2013
IRSHAD_1	3	Prostate cancer	Irshad et al., 2013
LARKIN	7	Prostate cancer	Larkin et al., 2012
LI	6	Prostate cancer	Li et al., 2019
LIU	167	Multiple cancers	Liu et al., 2007
LONG	12	Prostate cancer	Long et al., 2011
NAKAGAWA	17	Prostate cancer	Nakagawa et al., 2008
PENNEY	157	Prostate cancer	Penney et al., 2011
RAMASWAMY	16	Multiple cancers	Ramaswamy et al., 2003
REDDY	16	Prostate cancer	Reddy and Balk, 2006
ROSS-ADAMS	100	Prostate cancer	Ross-Adams et al., 2015
ROSS	6	Prostate cancer	Ross et al., 2012
SAAL	162	Multiple cancers	Saal et al., 2007
SHARMA	15	Prostate cancer	Sharma et al., 2013
SINGH	5	Prostate cancer	Singh et al., 2002
SONG	15	Prostate cancer	Song et al., 2019
STEPHENSON	10	Prostate cancer	Stephenson et al., 2005
TALANTOV	3	Prostate cancer	Talantov et al., 2010
TANDEFELT	36	Prostate cancer	Tandefelt et al., 2013
TRUE	86	Prostate cancer	True et al., 2006
WANG	43	Prostate cancer	Wang et al., 2017
WU	29	Prostate cancer	Wu et al., 2013
YU	14	Multiple cancers	Yu et al., 2007

Number of genes corresponds to the number of genes in a signature after conversion from HGNC gene symbols to Entrez gene IDs.*

The prediction results are assessed using the *p*-values obtained from the survival analysis. We call a random gene set with a significant *p*-value, a *surrogate gene set*.

2.5. Unsupervised Classification

The patient samples are categorized using the PC1 stratification method, which is based on a principal component analysis (PCA). Briefly, PCA is a dimension reduction technique (this involves reducing the size of the data set). The goal is to transform a large data set into a smaller ones having a lower dimensional representation. This method trades a little accuracy for simplicity, thus achieving interpretability as well as minimal loss of information (Lever et al., 2017). For performing the PC1 method, we use the R function "prcomp" to obtain the first principal component (PC1) of a signature. The patients are then divided into two groups according to the median of the PC1, i.e., a sample is either categorized as group -1 if the PC1 is below the median or as group $+1$ if the PC1 is above the median value. Hence, the PC1 method is used to classify (or group) the patients into two classes, whereas this separation depends on a signature gene set. This approach has been previously used (see, e.g., Venet et al., 2011).

Formally, our analysis is based on a gene expression matrix of the form $X \in \mathbb{R}^m \times \mathbb{R}^n$, where m is the number of genes and n is the number of samples. Importantly, here m corresponds to the number of genes in a particular signature gene signature and not to all genes that are available in a data set.

2.6. Survival Analysis

For evaluating the prognostic value of gene sets, we conduct a survival analysis. Specifically, we estimate a Kaplan–Meier survival curve for each patient group and compare different groups with the Mantel–Haenszel test (Emmert-Streib and Dehmer, 2019). Hence, each comparison is characterized by a *p*-value resulting from a statistical hypothesis test. For the survival analysis, we use the progression-free interval as endpoint.

We would like to remark that due to the fact that the PC1 method provides a categorization of the patients, the resulting survival analysis depends on the gene set used for obtaining the first principal component of the signature.

2.7. Measuring of Biological Meaning

In order to have a well-defined meaning of the term "biological meaning," we use information from the GO (Ashburner et al., 2000). Specifically, GO defines the biological meaning of a gene by a list of GO-terms associated with this gene. For a list of genes, the biological meaning of this set can be defined by the union of the sets of GO-terms of the individual genes. For instance, given three genes, g_1, g_2, g_3, with associate GO-terms the biological meaning (M) of these genes is given by

$$M(g_1) = \{GO_1(1), GO_1(2), \ldots GO_1(m)\} \quad (2)$$

$$M(g_2) = \{GO_2(1), GO_2(2), \ldots GO_2(n)\} \quad (3)$$

$$M(g_3) = \{GO_3(1), GO_3(2), \ldots GO_3(o)\} \quad (4)$$

with $m, n, o \in \mathbb{N}$. Here, the GO-terms are from a category, e.g., BP. Similarly, the biological meaning of the set of genes $\{g_1, g_2, g_3\}$ is given by

$$M(\{g_1, g_2, g_3\}) = M(g_1) \cup M(g_2) \cup M(g_3) \quad (5)$$

whereas $\cup$ is the union of the individual sets. Hence, the biological meaning of $\{g_1, g_2, g_3\}$ is given by the set of GO-terms $M(\{g_1, g_2, g_3\})$.

From this follows that, e.g., the similarity of two sets of genes, $\{g_1, g_2, g_3\}$ and $\{g_4, g_5, g_6\}$, is zero if

$$M(\{g_1, g_2, g_3\}) \cap M(\{g_4, g_5, g_6\}) = \emptyset. \quad (6)$$

Importantly, our GRP defined above constructs random gene sets (RGS) with this property, i.e.,

$$M(RGS) \cap M(BM) = \emptyset \quad (7)$$

with RGS a set of random genes and BM a set of biomarker genes.

3. RESULTS

In this section, we present the results of our analysis. First, we study published prognostic biomarkers of PCa individually and comparatively. Then we study random gene set and show results for prognostic outcome.

3.1. Prognostic Biomarkers of Prostate Cancer

3.1.1. Size of Biomarker Sets and GO-Terms in Signatures

In **Table 1**, we show an alphabetically ordered overview of all 32 prognostic BM signatures included in our analysis. The smallest signature is from Cheville consisting of 2 genes only, whereas the signature from Liu is the largest containing 167 genes. Interestingly, there are some signatures that have the same number of genes. Specifically, the signatures of Irshad_1 and Talantov have 3 genes, the signatures of Li and Ross have 6 genes, the signatures of Chen_cc and Larkin have 7 genes, the signatures of Agell, Bismar, and Long have 12 genes, the signatures of Sharma and Song have 15 genes, and the signatures of Bibikova, Ramaswamy, and Reddy have 16 genes in their BM sets. An overall summary of the size distributions of all BM signatures is shown in **Figure 1A**.

In **Figure 1B**, we show information about the GO-terms associated with the genes in the signatures. The three colors correspond to the three GO categories: BP shown in cyan, molecular function (MF) shown in red, and cellular component (CC) shown in yellow. For each of these three categories, we show the absolute number of GO-terms in each study. Overall, from **Figure 1B** one can see that the present GO-terms in the signatures differ significantly from each other. That means some signatures are very specific because they contain only a very small number of different GO-terms, e.g., the signatures from Talantov, Cheville and Li, while others are rather generic containing many GO-terms, e.g., Penney, Liu and Saal. For GO-terms of BP (cyan), this variation is particularly large.

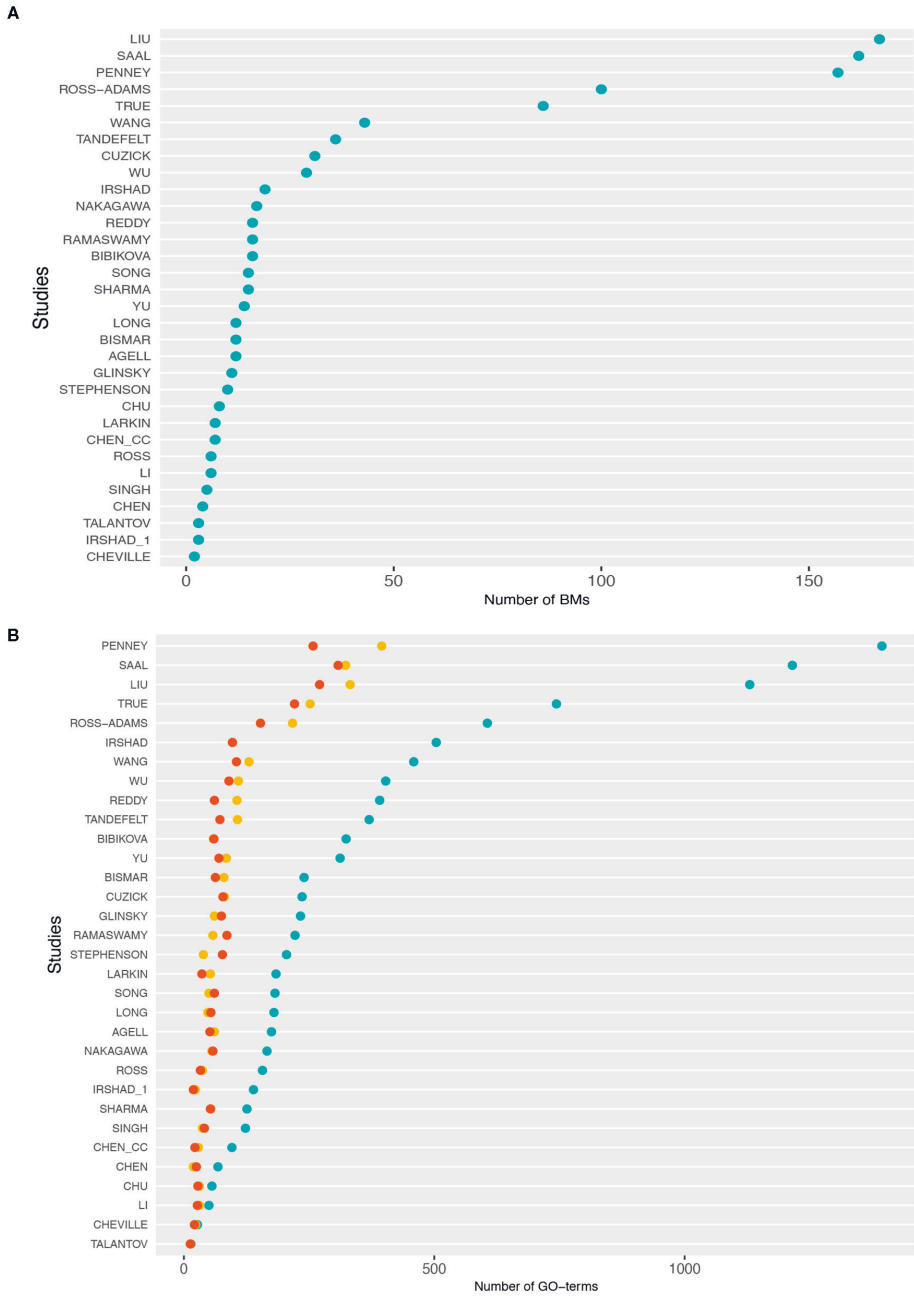


FIGURE 1 | (A) Distributions of the number of biomarker genes in each study. **(B)** Number of gene ontology (GO)-terms associated with the signature genes in each study. The cyan points correspond to BP, the red points to molecular function (MF), and yellow points to cellular component (CC).

3.1.2. Pairwise Similarity of Signatures

In order to study differences between the 32 signatures, we conduct a pairwise comparison of these BM sets. Specifically, we study two different types of overlap. We study (i) the number of common genes and (ii) the number of common GO-terms among pairs of signatures. Formally, we define these two overlap measures as follows. Let S_i and S_j be two signature sets consisting either of genes or GO-terms. Then we find the percentage $z_i \in [0, 1]$ of common elements in S_i that are also present in S_j by

$$x_i = S_i \cap S_j \tag{8}$$

$$z_i = \frac{|x_i|}{|S_i|} \tag{9}$$

Here, z_i can assume values between zero and one and $|z|$ corresponds to the number of elements in set z. We would like to remark that the way we define the overlap is asymmetric, i.e., $z_i \neq z_j$ if $|S_i| \neq |S_j|$. That means the percentage of the overlap is taken with respect to the first signature set S_i on the right-hand side of Equation (8).

The two heatmaps in **Figures 2A,B** show the results of this analysis. From this analysis of the gene overlap, we find that the signatures of Chen_cc and Chu do not overlap with other signatures at all, i.e., both have a zero overlap with any other signature. This implies that the genes in the signatures of Chen_cc and Chu are unique concerning the genes in their corresponding BM sets. Every other BM signature has at least some overlap with another signature; see the last column in **Figure 2A** (red numbers) providing information about the number of signatures with a non-zero overlap.

The signature of Cheville, which has the smallest number of genes, has a gene overlap with the three signatures of Cuzick, Li, and Penney. Surprisingly, the signature of Liu, which contains the highest number of genes, has only genes in common with 9 other signatures. Irshad_1 is the only signature that completely overlaps with another signature (Irshad); however, we would like to note that both signatures are from the same study (Irshad et al., 2013). Finally, we find that the signature of Penney has the highest gene overlap with other signatures (it has genes in common with 20 signatures). From this analysis, we see that there is a wide range of behaviors for the gene overlap reaching from zero overlap (for Chen_cc and Chu) to an overlap with 20 signatures (for Penny) corresponding to an overlap with 64.5% ($= 20/31$) of all signatures. This implies that all signatures are unique to a certain extend because this percentage would be much higher.

In contrast to these findings, **Figure 2B** shows the overlap of GO-terms among the signatures. Again, the overlap between the signatures varies considerably. For instance, the signatures of Saal and Penney share the highest overlap with 490 GO-terms. Interestingly, all the signatures have a non-zero overlap in their biological meaning.

Importantly, a difference to the gene overlap (see **Figure 2A**) is that for a GO-term overlap, all signatures share at least one GO-term with 26 other signatures (see last column in **Figure 2B**) and most signatures (25) have at least one common GO-term with all other signatures. This implies that on a GO-term level, the signatures are much more similar to each other than on a gene level. This underlines the importance of a systems-view on PCa.

3.2. Prediction Abilities of Random Gene Signatures

Next, we systematically investigated the prognostic prediction capabilities of the 32 BM signatures and random gene sets. We begin by systematically removing BM signature genes from the available gene expression gene pool. Subsequently, we also omit hierarchically genes that share a biological meaning with the respective published signatures. We randomly sample 1,000 set of the same size as the BM signature from the gene set left to create random gene signature. The results are as follows:

The outcome of the study is given in three parts. First, from the gene pool, we systematically remove the published signatures and the genes that share a similar biological meaning with them and compute the outcome association. Next, we correct the obtained p-values by conservative Bonferroni correction and report the results. And finally, the analysis is repeated by omitting the proliferation genes from the gene pool, we correct the p-values by conservative Bonferroni correction, and present the results.

3.2.1. GDC Cohort a Data

The results for the GDC cohort A data are shown in **Figure 3**. The light/dark red points represent the outcome of the published signatures (without any gene removal), whereas light red indicates significant results and dark red non-significant one. The blue colored distributions are the result of random gene sets, whereas the shaded cyan bars correspond to the lower third percentile of the distributions and the bold black points are the median values of these distributions. The blue vertical line corresponds to a significant level of $\alpha = 0.05$. We would like to note that the p-values are on a logarithmic scale (i.e., $\log_{10}$).

First, from **Figure 3** we observed that not all published signatures (red points) lead to significant results. In order to highlight this, we show significant results by points in light red, whereas non-significant results are shown in dark red. A possible reason for this observation is that our analysis uses a different data set than the original studies and, hence, the observed results indicate to the well-known instability of biomarkers (lack of robustness) (Drier and Domany, 2011). Specifically, for our analysis 24 of the 32 biomarker signatures are significant and the remaining published signatures lack robustness for the independent validation data set.

Figure 3A shows results without a Bonferroni correction. This analysis is similar to the study by Venet et al. (2011), which also did not use a multiple testing correction even though many comparisons were conducted. Interestingly, in **Figure 3A** all lower third percentiles (cyan shaded bars) are significant. That means for all random gene sets we find at least 3% of these to be significant. When compared to the published signatures (red points), the lower third percentile of random gene sets outperform even 26 signatures. Five published signatures performed as well as the lower third percentile of random sets, or the random sets slightly outperformed them. Only one signature (Song) achieves a more significant outcome than the lower third percentile of the random gene sets. Two signatures Ross and

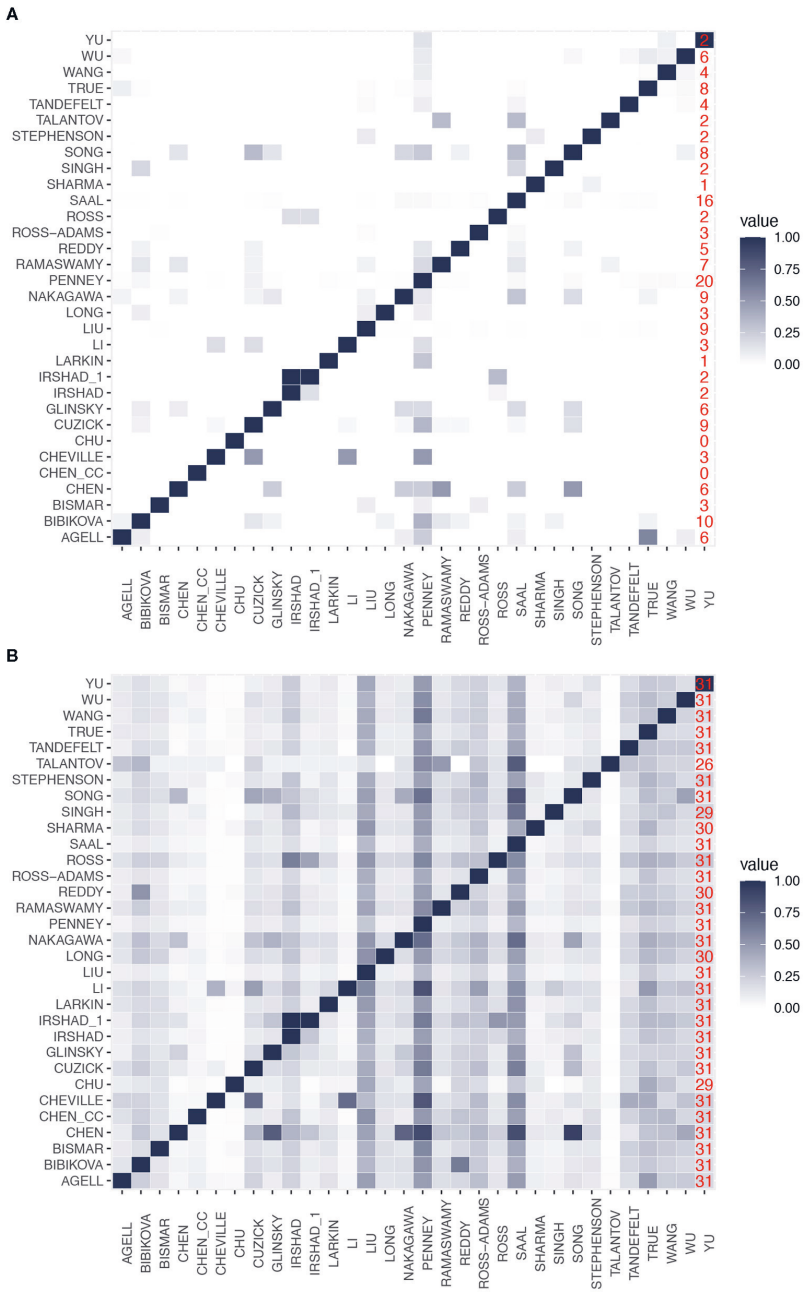


FIGURE 2 | Pairwise overlap between prognostic biomarkers. **(A)** Overlap in the number of genes. **(B)** Overlap in the number of gene ontology (GO)-terms. The last column in both heatmaps (red numbers) gives the number of signatures with a non-zero overlap.

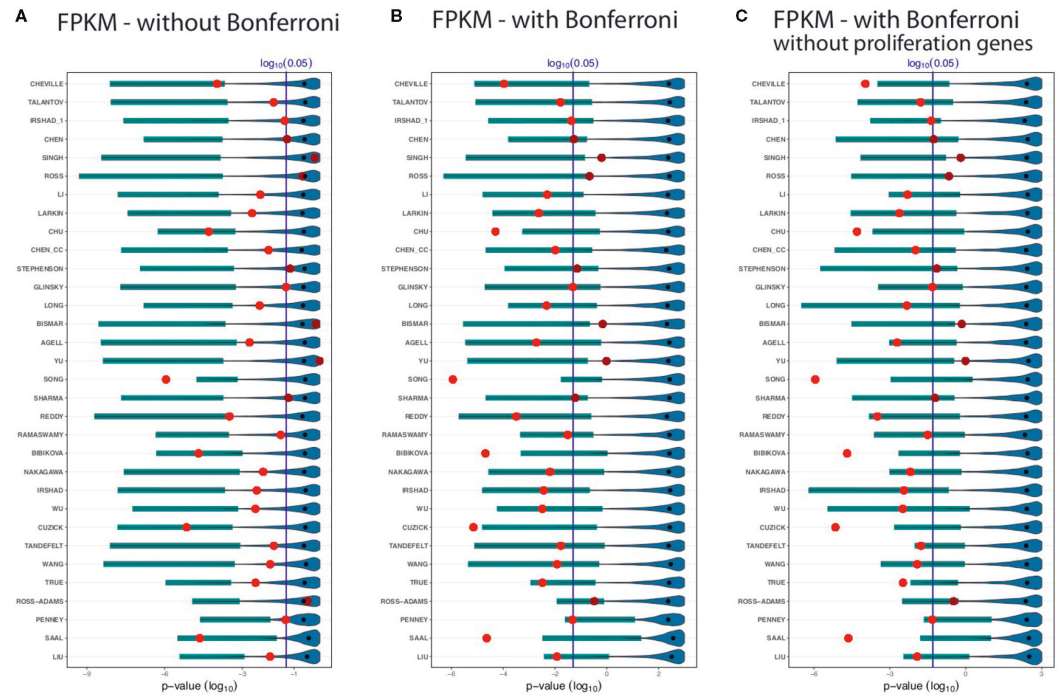


FIGURE 3 | Shown are the prognostic prediction results of random gene sets for 32 signatures using the GDC cohort A data. **(A)** Results for uncorrected *p*-values. **(B)** Bonferroni corrected *p*-values. **(C)** Proliferation genes are removed and the *p*-values are Bonferroni corrected. The significance level is $\alpha = 0.05$ (vertical blue line) and the light/dark red points represent the outcome of the published signatures. The blue-colored distributions are the results of the random gene sets, whereas the shaded cyan bars correspond to the lower third percentile of the distributions and the bold black points are the median values.

Ross-Adams perform as worse as the median of the random sets and three signatures (Singh, Bismar, and Yu) perform even worse than the median of the random gene sets. The median of the random sets (bold black points) are all non-significant.

In **Figure 3B**, we repeated the analysis applying a conservative Bonferroni correction. With a Bonferroni correction, four signatures (Singh, Ross, Bismar, and Yu) performed worse than the lower third percentile of random signatures. Likewise, five published signatures, Chu, Song, Bibikova, Cuzick, and Saal, outperformed the random signatures. As one can see from **Figure 3B**, not all the lower third percentile are significant. However, for all random signatures (such as Penney and Liu), we find at least some significant random signatures. Interestingly, many smaller random signatures perform better in comparison to larger ones. For instance, Cheville, Talantov, Irshad_1, Chen, Singh, etc., all performed better than the top 5 largest signatures (True, Ross-Adams, Penney, Saal, and Liu).

In a previous breast cancer study (Goh and Wong, 2018), it has been found that the removal of proliferation genes from random signatures leads to diminishing results of the prognostic performance of random signatures. In order to study this effect, we removed additionally all proliferation genes from the gene pool and repeated our analysis with a Bonferroni correction. The results of this are shown in **Figure 3C**. Qualitatively, the results in **Figures 3B,C** are similar. Overall, for all results in **Figures 3A–C**, one can see that for all random signatures there are at least some that are statistically significant. We would like to emphasize that all random gene sets share *per construction* no biological meaning with the published signatures yet can perform prognosis as well as the BM signatures or better.

3.2.2. GDC Cohort B Data

In order to study the influence of the data processing, we repeat our analysis for the GDC cohort B data. The results of this analysis are shown in **Figures 4A–C**. In these figures, there are in addition to the dark and light red points, light green points indicate the BM signatures. These correspond to significant BM signatures, whereas the median values of the random gene sets (black points) are non-significant.

Again, we observe that not all BM signatures lead to a significant outcome. Specifically, we find 22 of the 32 signatures to be significant (**Figure 4**). Interestingly, we find also non-robust

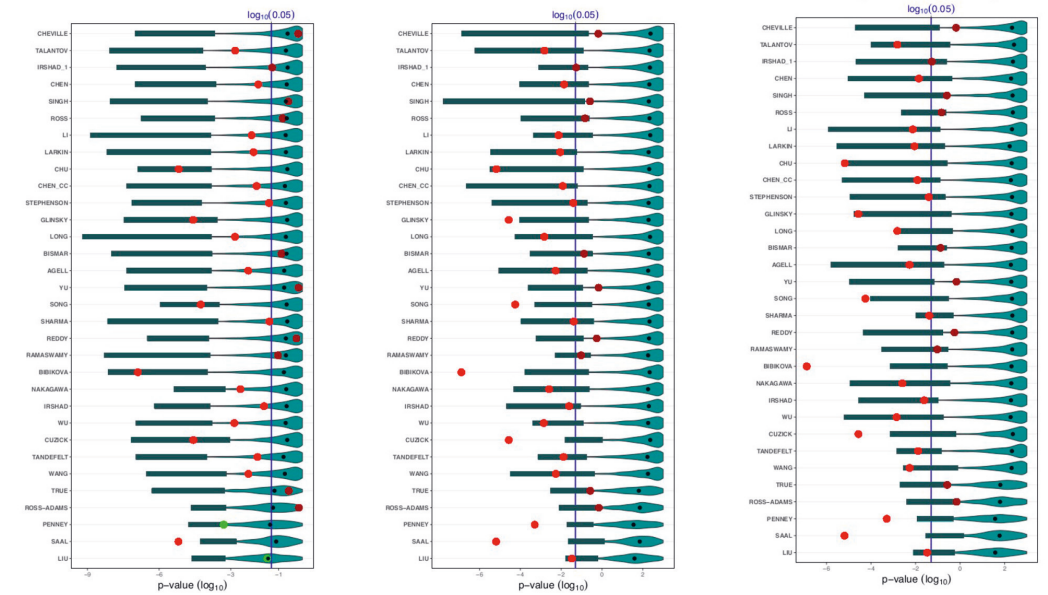


FIGURE 4 | Shown are the prognostic prediction results of random gene sets for 32 signatures using the GDC cohort B data set. **(A)** Results for uncorrected *p*-values. **(B)** Bonferroni corrected *p*-values. **(C)** Proliferation genes are removed and the *p*-values are Bonferroni corrected. The significance level is $\alpha = 0.05$ (vertical blue line) and the light/dark red points represent the outcome of the published signatures. The blue colored distributions are the results of the random gene sets, whereas the shaded cyan bars correspond to the lower third percentile of the distributions and the bold black points are the median values.

results. For instance, Cheville, Irshad_1, Reddy, Ramaswamy, and True failed to predict the outcome in the GDC cohort B data set, but these signatures were significant for the GDC cohort A data (see **Figure 3**). Similarly, Chen, Stephenson, and Sharma are significant for GDC cohort B (see **Figure 4**) but not GDC cohort A (see **Figure 3**).

Also for the distributions of the results for the random gene sets, we observe very similar results as for the GDC cohort A data in **Figure 3**. Hence, overall, the results in **Figures 4A–C** confirm our analysis, which means there are always random gene sets leading to significant results.

4. DISCUSSION

Our hypothesis for the present study was that prognostic signatures of prostate cancer are lacking a sensible biological meaning. In order to investigate this, we used a GRP introduced in Manjang et al. (2021). This GRP allows to systematically construct random gene sets by omitting all biological similarities between published signatures and the genes in a gene pool from which random gene sets are drawn. These random gene sets are not assigned any particular (biological) meaning or role. Importantly, such random gene sets do not necessarily have predictive capabilities as assessed by predicting progression-free survival as outcome variable. For this reason, we distinguish between random gene sets that are predictive (indicated by a significant *p*-value from a survival analysis) and non-predictive by calling the former ones surrogate gene sets. A published BM signature (see **Table 1**), on the other hand, is a gene set that is obtained in a targeted and non-random manner indicative of disease progression.

For testing our hypothesis, we studied 32 published BM signatures of prostate cancer from the literature (see **Table 1**). As a result, for all studied 32 signatures we found random gene sets with better or similar prognostic capabilities but no overlap in the biological meaning. In order to see if the preprocessing of the data has any effect on this, we extended our analysis by examining the effect of different data processing techniques. Regarding this, we conducted further analysis by using a data set with different processing methods applied to the raw data leading to a data set we call GDC cohort B. As a result from this analysis, we found no systematic influence of a particular data processing technique on the surrogate gene sets or the overall results (see **Figures 3**, **4**). Finally, we also removed proliferation

genes (for both data sets, i.e., GDC cohort A and GDC cohort B) and found also for this setting no difference in our results (see **Figures 3C**, **4C**).

As a conclusion from all these analyses, we can infer that any biological rationale provided for selecting the genes in the published gene signatures, as shown in **Table 1**, is anecdotal. This is taking into account the meaning of random gene sets arising from the GRP because the used GRP eliminates the risk of accidentally selecting genes for a random gene set that have the same biological meaning as the published gene signatures. Consequently, due to the discovery of surrogate gene sets with the same predictive capability but a completely distinct biological interpretation, as a result of the zero overlap in the GO-terms of the genes involved, any biological significance attributed to such BM signatures is required.

Interestingly, a similar interpretation has been found in a breast cancer study by Manjang et al. (2021). They showed that when all signs of the biological meaning of the BM signature genes are removed, surrogate gene sets can be determined among the remaining random gene sets with similar prognostic predictive capabilities but with contrasting biological meaning. Therefore, the research findings indicated that with regard to disease etiology, none of the studied signatures have a plausible biological interpretation or significance. The study concluded that prognostic signatures are black-box models that can yield accurate predictions of breast cancer outcome but with no benefit for disclosing causal, biological relations. Furthermore, this study also noted a relationship between the predictive accuracy and the size of the random gene sets by showing that the accuracy is higher for larger gene sets. It is interesting to note that in the current study, we could not establish this relationship. A possible explanation for this may be the relatively small size of published BM signatures of prostate cancer, which are all smaller than 200 genes (see **Figure 1**). In contrast, the breast cancer signatures studied in Manjang et al. (2021) are much larger in average reaching up to 1345 genes.

It is important to note that a similar study for breast cancer by Venet et al. (2011) has been unable to arrive at this conclusion since no GRP was used. As a consequence, BM signatures as well as genes from associated BP were not removed leaving the possibility to inadvertently select random genes with a common biological meaning as the original BM signatures because these genes belong to the same BPs as indicated by common GO-terms in the domain BP. Another statement by Venet et al. (2011) is that *most random signatures are significantly associated with prognostic outcome*. With respect to prostate cancer, this holds only for the random gene sets of Penney and Liu (see **Figure 4A**) because 50% of the surrogate gene sets are significant as indicated by the median values of the distributions (black points in **Figure 4**). However, generally, this assertion is not valid and only applies to some signatures.

To date, many studies investigated prognostic signatures of prostate cancer. For example, Bibikova et al. (2007) used a 16-gene expression signature to predict the prognosis of prostate cancer. They complemented their results by a discussion of the functional annotation of these genes, which were involved in proliferation, cell cycle, differentiation, signal transduction and basic metabolism. Similarly, the studies by Saal et al. (2007), Sharma et al. (2013), and Song et al. (2019) argued that the biological importance of their prognostic signatures is based on the role of PI3K pathway signaling, altered signaling, P53 signaling and cell cycle process pathway respectively. In this paper, we studied those and other prognostic signatures of prostate cancer. Our results, however, demonstrate that such biological interpretations do not offer a causal explanation for the fundamental biology of prostate cancer since we can always find surrogate gene sets with no biological relationship to those signatures but similar or better prognostic prediction capabilities.

Considering that prostate cancer and breast cancer are two considerably different diseases yet our results demonstrate a similarity in the lack of biological meaning of both cancers one may wonder if there is a common factor giving raise to these findings. This is very difficult to answer, however, one common factor that comes to mind are the hallmarks of cancer (Hanahan and Weinberg, 2000). Specifically, the study by Hanahan and Weinberg (2000) highlighted six hallmarks of cancer (self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis), which are shared by all types of human cancers. Later this has been extended by four further hallmarks (deregulating cellular energetics, avoiding immune destruction, genome instability and mutation, tumor-promoting inflammation) (Hanahan and Weinberg, 2011). If our findings are actually related to the ten hallmarks of cancer is currently unclear. However, it seems not implausible to assume that there might be a connection because the hallmarks state that cancer is a system disease involving a multitude of pathways. We want to add that these pathways do not work in isolation but are connected among each other by intricate regulatory networks (Emmert-Streib et al., 2014).

On a technical note, we would like to remark that there could be other metrics for evaluating the prediction capabilities of random gene sets other than p-values. For instance, one could use information from pathology about disease states, which allow to use error measures for binary classifications. While this establishes sensible metrics, e.g., F-score or AUROC, such measures do not directly utilize survival information about the progression of patients. Instead, this is the strength of survival analysis comparing survival curves quantitatively. Hence, a regression framework, as provided by survival analysis (Kleinbaum and Klein, 2005), seems to be favorable over a classification framework allowing a more nuanced evaluation.

Finally, we would like to note that our study has similarities to recent investigations in Explainable Artificial Intelligence (XAI) (Xu et al., 2019; Emmert-Streib et al., 2020). Specifically, XAI explores the dichotomy of predictive and descriptive models (Emmert-Streib and Dehmer, 2021) in AI and aims to establish mechanisms for making predictive models also explainable in a sense that this can enhance our understanding of a system under investigation. On a wider scope, this discussion has a long history in the statistics community and refers to the distinction of black-box models and causal models (Holland, 1986; Breiman, 2001). Our study shows that prognostic biomarkers of prostate

cancer allow sensible predictions for cancer progression but do not establish a causal understanding with respect to the biological meaning of such prognostic signatures. Here, it is important to extend the considerations to the proposed gene selection mechanisms used by studies identifying prognostic signatures (see **Table 1**). Overall, such models have a predictive utility, e.g., for applications in the clinical practice but no biological utility for enhancing our understanding of cancer biology.

5. CONCLUSION

In this paper, we scrutinized the biological meaning of prognostic signatures of prostate cancer. Our study utilized a GRP that results in random gene sets without any overlap in the biological meaning with biomarker signatures yet a non-vanishing proportion of these random gene sets, called surrogate gene sets, achieve similar prediction results. Hence, our results demonstrate that none of the studied signatures of prostate cancer has a sensible biological interpretation with respect to disease etiology. To our knowledge, this is the first study providing such results for prognostic biomarkers of prostate cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://xenabrowser.net/datapages/.

AUTHOR CONTRIBUTIONS

FE-S conceived the study. KM performed the analysis. KM and FE-S analyzed the data and interpreted the results. All authors wrote the manuscript.

FUNDING

KM has been supported by a fellowship from the Center for Prostate Cancer, Tampere University. MD thanks the Austrian Science Funds for supporting this work (project P30031).

REFERENCES

Agell, L., Hernández, S., Nonell, L., Lorenzo, M., Puigdecanet, E., de Muga, S., et al. (2012). A 12-gene expression signature is associated with aggressive histological in prostate cancer: Sec14l1 and tceb1 genes are potential markers of progression. *Am. J. Pathol.* 181, 1585–1594. doi: 10.1016/j.ajpath.2012.08.005

Ashburner, M., Ball, C., Blake, J., Botstein, D., and Butler H., et al. (2000). Gene Ontology: tool for the unification of biology. The Gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bibikova, M., Chudin, E., Arsanjani, A., Zhou, L., Garcia, E. W., Modder, J., et al. (2007). Expression signatures that correlated with gleason score and relapse in prostate cancer. *Genomics* 89, 666–672. doi: 10.1016/j.ygeno.2007.02.005

Bismar, T. A., Demichelis, F., Riva, A., Kim, R., Varambally, S., He, L., et al. (2006). Defining aggressive prostate cancer using a 12-gene model. *Neoplasia* 8:59. doi: 10.1593/neo.05664

Breiman, L. (2001). Statistical modeling: the two cultures. *Stat. Sci.* 16, 199–231. doi: 10.1214/ss/1009213726

Chen, X., Wang, J., Peng, X., Liu, K., Zhang, C., Zeng, X., et al. (2020). Comprehensive analysis of biomarkers for prostate cancer based on weighted gene co-expression network analysis. *Medicine* 99:e19628. doi: 10.1097/MD.0000000000019628

Chen, X., Xu, S., McClelland, M., Rahmatpanah, F., Sawyers, A., Jia, Z., et al. (2012). An accurate prostate cancer prognosticator using a seven-gene signature plus gleason score and taking cell type heterogeneity into account. *PLoS ONE* 7:e45178. doi: 10.1371/journal.pone.0045178

Cheville, J. C., Karnes, R. J., Therneau, T. M., Kosari, F., Munz, J.-M., Tillmans, L., et al. (2008). Gene panel model predictive of outcome in men at high-risk of systemic progression and death from prostate cancer after radical retropubic prostatectomy. *J. Clin. Oncol.* 26:3930. doi: 10.1200/JCO.2007.15.6752

Chu, J., Li, N., and Gai, W. (2018). Identification of genes that predict the biochemical recurrence of prostate cancer. *Oncol. Lett.* 16, 3447–3452. doi: 10.3892/ol.2018.9106

Cuzick, J., Swanson, G. P., Fisher, G., Brothman, A. R., Berney, D. M., Reid, J. E., et al. (2011). Prognostic value of an rna expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *Lancet Oncol.* 12, 245–255. doi: 10.1016/S1470-2045(10)70295-3

Drier, Y., and Domany, E. (2011). Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? *PLoS ONE* 6:e17795. doi: 10.1371/journal.pone.0017795

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proce. Natl. Acad. Sci. U.S.A.* 103, 5923–5928. doi: 10.1073/pnas.0601231103

Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B., and Dehmer, M. (2014). The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Front. Genet.* 5:15. doi: 10.3389/fgene.2014.00015

Emmert-Streib, F., and Dehmer, M. (2019). Introduction to survival analysis in practice. *Mach. Learn. Knowl. Extract.* 1, 1013–1038. doi: 10.3390/make1030058

Emmert-Streib, F., and Dehmer, M. (2021). Data-driven computational social network science: predictive and inferential models for web-enabled scientific discoveries. *Front. Big Data* 4:591749. doi: 10.3389/fdata.2021.591749

Emmert-Streib, F., Yli-Harja, O., and Dehmer, M. (2020). Explainable artificial intelligence and machine learning: a reality rooted perspective. *WIREs Data Min. Knowl. Discov.* 10:e1368. doi: 10.1002/widm.1368

Gilhodes, J., Zemmour, C., Ajana, S., Martinez, A., Delord, J.-P., Leconte, E., et al. (2017). Comparison of variable selection methods for high-dimensional survival data with competing events. *Comput. Biol. Med.* 91, 159–167. doi: 10.1016/j.compbiomed.2017.10.021

Glinsky, G. V., Berezovska, O., Glinskii, A. B., et al. (2005). Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J. Clin. Invest.* 115, 1503–1521. doi: 10.1172/JCI23412

Goh, W. W. B., and Wong, L. (2018). Why breast cancer signatures are no better than random signatures explained. *Drug Discov. Today* 23, 1818–1823. doi: 10.1016/j.drudis.2018.05.036

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE* 6:e28210. doi: 10.1371/journal.pone.0028210

Holland, P. (1986). Statistics and causal inference. *J. Am. Stat. Assoc.* 81, 945–960. doi: 10.1080/01621459.1986.10478354

Irshad, S., Bansal, M., Castillo-Martin, M., Zheng, T., Aytes, A., Wenske, S., et al. (2013). A molecular signature predictive of indolent prostate cancer. *Sci. Transl. Med.* 5:202ra122. doi: 10.1126/scitranslmed.3006408

Kim, S.-Y. (2009). Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics* 10:147. doi: 10.1186/1471-2105-10-147

Kleinbaum, D. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text. Statistics for Biology and Health*. New York, NY: Springer.

Larkin, S., Holmes, S., Cree, I., Walker, T., Basketter, V., Bickers, B., et al. (2012). Identification of markers of prostate cancer progression using candidate gene expression. *Br. J. Cancer* 106, 157–165. doi: 10.1038/bjc.2011.490

Lever, J., Krzywinski, M., and Altman, N. (2017). Points of significance: principal component analysis. *Nat. Methods* 14, 641–642. doi: 10.1038/nmeth.4346

Li, F., Ji, J.-P., Xu, Y., and Liu, R.-L. (2019). Identification a novel set of 6 differential expressed genes in prostate cancer that can potentially predict biochemical recurrence after curative surgery. *Clin. Transl. Oncol.* 21, 1067–1075. doi: 10.1007/s12094-018-02029-z

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416. doi: 10.1016/j.cell.2018.02.052

Liu, R., Wang, X., Chen, G. Y., Dalerba, P., Gurney, A., Hoey, T., et al. (2007). The prognostic role of a gene signature from tumorigenic breast-cancer cells. *New Engl. J. Med.* 356, 217–226. doi: 10.1056/NEJMoa063994

Long, Q., Johnson, B. A., Osunkoya, A. O., Lai, Y.-H., Zhou, W., Abramovitz, M., et al. (2011). Protein-coding and microrna biomarkers of recurrence of prostate cancer following radical prostatectomy. *Am. J. Pathol.* 179, 46–54. doi: 10.1016/j.ajpath.2011.03.008

Manjang, K., Tripathi, S., Yli-Harja, O., Dehmer, M., and Emmert-Streib, F. (2020). Graph-based exploitation of gene ontology using goxplorer for scrutinizing biological significance. *Sci. Rep.* 10, 1–16. doi: 10.1038/s41598-020-73326-3

Manjang, K., Tripathi, S., Yli-Harja, O., Dehmer, M., Glazko, G., and Emmert-Streib, F. (2021). Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. *Sci. Rep.* 11, 1–18. doi: 10.1038/s41598-020-79375-y

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365, 488–492. doi: 10.1016/S0140-6736(05)17866-0

Nakagawa, T., Kollmeyer, T. M., Morlan, B. W., Anderson, S. K., Bergstralh, E. J., Davis, B. J., et al. (2008). A tissue biomarker panel predicting systemic progression after psa recurrence post-definitive prostate cancer therapy. *PLoS ONE* 3:e2318. doi: 10.1371/journal.pone.0002318

Penney, K. L., Sinnott, J. A., Fall, K., Pawitan, Y., Hoshida, Y., Kraft, P., et al. (2011). mrna expression signature of gleason grade predicts lethal prostate cancer. *J. Clin. Oncol.* 29:2391. doi: 10.1200/JCO.2010.32.6421

Ramaswamy, S., Ross, K. N., Lander, E. S., and Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* 33, 49–54. doi: 10.1038/ng1060

Reddy, G. K., and Balk, S. P. (2006). Clinical utility of microarray-derived genetic signatures in predicting outcomes in prostate cancer. *Clin. Genitourin. cancer* 5, 187–189. doi: 10.3816/CGC.2006.n.035

Ross, R. W., Galsky, M. D., Scher, H. I., Magidson, J., Wassmann, K., Lee, G.-S. M., et al. (2012). A whole-blood rna transcript-based prognostic model in men with castration-resistant prostate cancer: a prospective study. *Lancet Oncol.* 13, 1105–1113. doi: 10.1016/S1470-2045(12)70263-2

Ross-Adams, H., Lamb, A., Dunning, M., Halim, S., Lindberg, J., Massie, C., et al. (2015). Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study. *EBioMedicine* 2, 1133–1144. doi: 10.1016/j.ebiom.2015.07.017

Saal, L. H., Johansson, P., Holm, K., Gruvberger-Saal, S. K., She, Q.-B., Maurer, M., et al. (2007). Poor prognosis in carcinoma is associated with a gene expression signature of aberrant pten tumor suppressor pathway activity. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7564–7569. doi: 10.1073/pnas.0702507104

Sharma, N. L., Massie, C. E., Ramos-Montoya, A., Zecchini, V., Scott, H. E., Lamb, A. D., et al. (2013). The androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. *Cancer Cell* 23, 35–47. doi: 10.1016/j.ccr.2012.11.010

Siegel, R., Miller, K., and Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 203–209. doi: 10.1016/S1535-6108(02)00030-2

Song, Z., Huang, Y., Zhao, Y., Ruan, H., Yang, H., Cao, Q., et al. (2019). The identification of potential biomarkers and biological pathways in prostate cancer. *J. Cancer* 10:1398. doi: 10.7150/jca.29571

Stephenson, A. J., Smith, A., Kattan, M. W., Satagopan, J., Reuter, V. E., Scardino, P. T., et al. (2005). Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer* 104, 290–298. doi: 10.1002/cncr.21157

Talantov, D., Jatkoe, T. A., Böhm, M., Zhang, Y., Ferguson, A. M., Stricker, P. D., et al. (2010). Gene based prediction of clinically localized prostate cancer progression after radical prostatectomy. *J. Urol.* 184, 1521–1528. doi: 10.1016/j.juro.2010.05.084

Tandefelt, D. G., Boormans, J. L., van der Korput, H. A., Jenster, G. W., and Trapman, J. (2013). A 36-gene signature predicts clinical progression in a subgroup of erg-positive prostate cancers. *Eur. Urol.* 64, 941–950. doi: 10.1016/j.eururo.2013.02.039

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

True, L., Coleman, I., Hawley, S., Huang, C.-Y., Gifford, D., Coleman, R., et al. (2006). A molecular correlate to the gleason grading system for prostate adenocarcinoma. *Proc. Natl. Acad. Sci. U.S.A.* 103, 10991–10996. doi: 10.1073/pnas.0603678103

Venet, D., Dumont, J. E., and Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* 7:e1002240. doi: 10.1371/journal.pcbi.1002240

Wang, L.-Y., Cui, J.-J., Zhu, T., Shao, W.-H., Zhao, Y., Wang, S., et al. (2017). Biomarkers identified for prostate cancer patients through genome-scale screening. *Oncotarget* 8:92055. doi: 10.18632/oncotarget.20739

Wu, C.-L., Schroeder, B. E., Ma, X.-J., Cutie, C. J., Wu, S., Salunga, R., et al. (2013). Development and validation of a 32-gene prognostic index for prostate cancer progression. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6121–6126. doi: 10.1073/pnas.1215870110

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (2019). "Explainable AI: a brief survey on history, research areas, approaches and challenges," in *CCF International Conference on Natural Language Processing and Chinese Computing* (Dunhuang: Springer), 563–574.

Yu, J., Yu, J., Rhodes, D. R., Tomlins, S. A., Cao, X., Chen, G., et al. (2007). A polycomb repression signature in metastatic prostate cancer predicts cancer outcome. *Cancer Res.* 67, 10657–10663. doi: 10.1158/0008-5472.CAN-07-2498

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.