

Deciphering gene expression patterns using large-scale transcriptomic data and its applications

Shunjie Chen¹, Pei Wang^{1,2,*}, Haiping Guo¹, Yujie Zhang¹

¹School of Mathematics and Statistics, Henan University, Jinming Avenue, 475004, Kaifeng, China

²Henan Engineering Research Center for Industrial Internet of Things, Henan University, Mingli Road, 450046, Zhengzhou, China

*Corresponding author: wangpei@henu.edu.cn, wp0307@126.com

Abstract

Gene expression varies stochastically across genders, racial groups, and health statuses. Deciphering these patterns is crucial for identifying informative genes, classifying samples, and understanding diseases like cancer. This study analyzes 11,252 bulk RNA-seq samples to explore expression patterns of 19,156 genes, including 10,512 cancer tissue samples and 740 normal samples. Additionally, 4,884 single-cell RNA-seq samples are examined. Statistical analysis using 16 probability distributions shows that normal samples display a wider range of distributions compared to cancer samples. Cancer samples tend to favor asymmetric distributions such as generalized extreme value, logarithmic normal, and Gaussian mixture distributions. In contrast, certain genes in normal samples exhibit symmetric distributions. Remarkably, more than 95.5% of genes exhibit non-normal distributions, which challenges traditional assumptions. Furthermore, distributions differ significantly between bulk and single-cell RNA-seq data. Many cancer driver genes exhibit distinct distribution patterns across sample types, suggesting potential for gene selection and classification based on distribution characteristics. A novel skewness-based metric is proposed to quantify distribution variation across datasets, showing genes with significant skewness differences have biological relevance. Finally, an improved naïve Bayes method incorporating gene-specific distributions demonstrates superior performance in simulations over traditional methods. This work enhances understanding of gene expression and its application in omics-based gene selection and sample classification.

Keywords: omics data; gene expression distribution; gene selection; skewness; naïve Bayes; sample classification

Introduction

Gene expression is a stochastic process that varies over time and across genders, racial groups, and health statuses [1–3]. However, previous investigations have primarily relied on limited snapshot data and assumed normal distributions [4–6]. With advancements in high-throughput sequencing technologies [7, 8], the accumulation of large-scale omics data presents new opportunities to elucidate gene distributions [9], identify informative genes [10–14], classify samples [15, 16], and explore complex diseases such as cancer [17, 18]. Yet, prior studies often focused on specific cancer tissues, adjacent normal tissues, and overlooked racial and gender differences, potentially biased by small sample sizes [19–30]. These challenges underscore the need to decode gene expression distributions and explore their applications through integrated analysis of large-scale omics data.

During the past decades, numerous methods have been developed to explore cancers using omics data. Common approaches include penalized regression models like LASSO [31], elastic net [32], and group LASSO [33], as well as Bayesian algorithms such as naïve Bayes (NB) [34] and its extensions [35, 36]. These methods often assume that gene expression data follow normal distributions [6], which facilitates theoretical analysis and ensures desired statistical properties [4–6]. However, based on the analysis of acute myeloid leukemia, ovarian cancer, and glioblastoma

multiforme datasets, Torrente *et al.* [9] found that most genes do not conform to normal distributions. They identified the optimal distribution for each gene as the one yielding the maximum *P*-value from the Kolmogorov–Smirnov (KS) test [37]. However, the KS test merely assesses distribution fit and does not measure goodness of fit. Marko *et al.* [38] further observed that genes in cancer tissues favored complex, heavy-tailed distributions with significant skewness and kurtosis, although their findings were based on limited data [38]. It remains unclear whether these observations hold across large-scale omics datasets. Recently, Zhang *et al.* [39] introduced a statistical method called IDEAS, which summarizes gene expression distributions within individuals and evaluates differences between groups. IDEAS heavily relies on denoising techniques and considers only simple distributions. With the growing availability of omics data, it is both necessary and intriguing to decipher gene expression distributions comprehensively using large-scale omics data.

Motivated by the aforementioned challenges, the emergence of an increasing number of large-scale omics datasets now presents us with unprecedented opportunities to delve deeply into the distributional characteristics of gene expression. Building on these discoveries, this study will leverage large-scale transcriptomic data to conduct a more in-depth analysis of gene expression distributions, aiming to fill gaps in the existing literature and

Received: August 2, 2024. **Revised:** October 7, 2024. **Accepted:** October 31, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

offer new perspectives on understanding the variations in gene expression across different populations and conditions. The contributions of this paper are primarily reflected in the following three aspects:

- Large-scale transcriptomic data are collected, and rigorous procedures are designed to decipher gene distribution patterns. Massive bulk RNA-seq data from TCGA [40] are utilized, categorized by clinical information. For each gene, its distribution is fitted with 16 typical probability distributions (Table 1). In contrast to prior work [9], the Bayesian Information Criterion (BIC) is employed to identify the potentially optimal distribution for each gene. Subsequently, the KS test is conducted to statistically evaluate whether the gene adheres to this distribution. The study addresses issues related to sample heterogeneity and elucidate reasons for the paucity of genes following normal distributions. Additionally, by focusing on key genes in breast cancer, the paper also explores differences in gene distribution patterns between bulk RNA-seq and single-cell RNA-seq (scRNA-seq) data.
- A novel metric, termed Skewness Ratio (SR), is introduced to quantify the skewness variation of genes between two datasets. This metric focuses on identifying genes that exhibit skewness differences in their expression distributions between two datasets, providing complementary insights to DESeq2 and edgeR [41, 42].
- An Improved Naïve Bayes (INB) classifier that integrates gene-specific distributions is developed. In the INB classifier, the optimal distribution fitted for each gene in a dataset is utilized to estimate the posterior probabilities of samples. Subsequently, samples are classified based on these posterior probabilities. While the INB model maintains the independence assumption among genes and incurs a slightly higher computational cost, it demonstrates superior performance in handling categorical data with diverse distributions. This advantage is particularly pronounced when dealing with non-normally distributed data, leading to more accurate classification results than traditional methods.

The schematic diagram depicting the main contents of this paper is shown in Fig. 1. The key content involves fitting gene-specific distributions, which are then utilized to identify informative genes and classify samples. This approach circumvents the assumption of a universally predefined distribution for all genes in a dataset, particularly challenging the normal distribution assumption. The proposed methods enable in-depth exploration of DEGs and enhance sample classification across diverse datasets.

Materials and methods

Data sources

Based on publicly available TCGA [40] database, bulk RNA-seq data for 33 types of cancers are collected, which consists of 11,252 samples, including 10,512 samples for cancer tissues and 740 samples for normal tissues. Gene expression levels were quantified using Transcripts Per Million (TPM) [43]. To focus on reliably expressed genes, those with zero expression in over 95% of the samples are filtered out, resulting in a final set of 19,156 protein-coding genes. The samples can be further stratified based on clinical variables such as gender, health status, race, and cancer stage.

Three publicly available datasets (GSE202203, GSE138536, GSE202695) from the NCBI database are also explored. After

filtering out genes with more than 95% zero expression values, 19,644 genes across 3,207 samples from the bulk RNA-seq dataset GSE202203 are considered for breast cancers. GSE138536 and GSE202695 are scRNA-seq datasets. GSE138536 initially includes expression data for 16,408 genes across 1,902 cells, while GSE202695 covers 23,459 genes across 2,982 cells. Genes with zero expression across all samples are excluded, resulting in the consideration of 10,000 genes across 862 cancer cells in GSE138536, and 18,273 genes across 771 cancer cells in GSE202695.

To evaluate the classification performance of the proposed classifier in this study, datasets for multiple cancer types including breast cancer (BRCA), hepatocellular carcinoma (HCC), kidney cancer (KICA), lung adenocarcinoma with lung squamous cell carcinoma (LUAD_LUSC), and thyroid carcinoma with bladder urothelial carcinoma (THCA_BLCA) are examined. BRCA comprises 110 normal samples and 116 cancer samples; HCC includes 50 normal and 50 cancer samples; KICA encompasses 126 normal and 126 cancer samples; LUAD_LUSC consists of 541 samples for lung adenocarcinoma and 502 samples for lung squamous cell carcinoma; THCA_BLCA includes 513 samples for thyroid carcinoma and 412 samples for bladder urothelial carcinoma.

The Kolmogorov–Smirnov test

The nonparametric KS test [37] assesses whether an observation vector $X_i = (x_{1i}, x_{2i}, \dots, x_{ni})^T$ for the i 'th gene is drawn from a distribution $F(x)$. The test statistic is defined as:

$$D = \sup_x |F_n(x) - F(x)|. \quad (1)$$

Here, $F_n(x)$ is the empirical distribution function based on the observations $x_{1i}, x_{2i}, \dots, x_{ni}$, which is defined as:

$$F_n(x) = \frac{1}{n} \sum_{t=1}^n I_{(-\infty, x]}(x_{ti}),$$

where $I_{(-\infty, x]}(x_{ti})$ is an indicator function that equals 1 if $x_{ti} < x$, and 0 otherwise. Essentially, $F_n(x)$ gives the proportion of observations that are less than or equal to x . $F(x)$ is the cumulative distribution function of the theoretical distribution being tested against. The test statistic D (Fig. 1) represents the maximum vertical deviation between the empirical distribution function $F_n(x)$ and the theoretical distribution function $F(x)$. It quantifies how well the empirical data matches the theoretical distribution. In this study, if the KS test P -value is no less than 0.01, then X_i can be considered to follow the distribution $F(x)$.

Principles for fitting gene expression distributions

Gene expression distributions are fitted using Python's stats and sklearn.mixture packages [44, 45]. A total of 16 distributions are considered, encompassing normal, t , Pareto, double Weibull (dweibull), generalized extreme value (genextreme), Laplace, Cauchy, χ^2 , exponential (expon), exponential power (exponpow), gamma, beta, logarithmic normal (log-norm), logarithmic gamma (log-gamma), uniform, and Gaussian mixture (GM) distributions. Table 1 outlines the general forms of their probability density functions (PDFs), covering a broad spectrum of distributions (Supplementary material). During distribution fitting, parameters for GM and dweibull distributions are estimated using the Expectation-Maximization (EM) algorithm, while parameters for

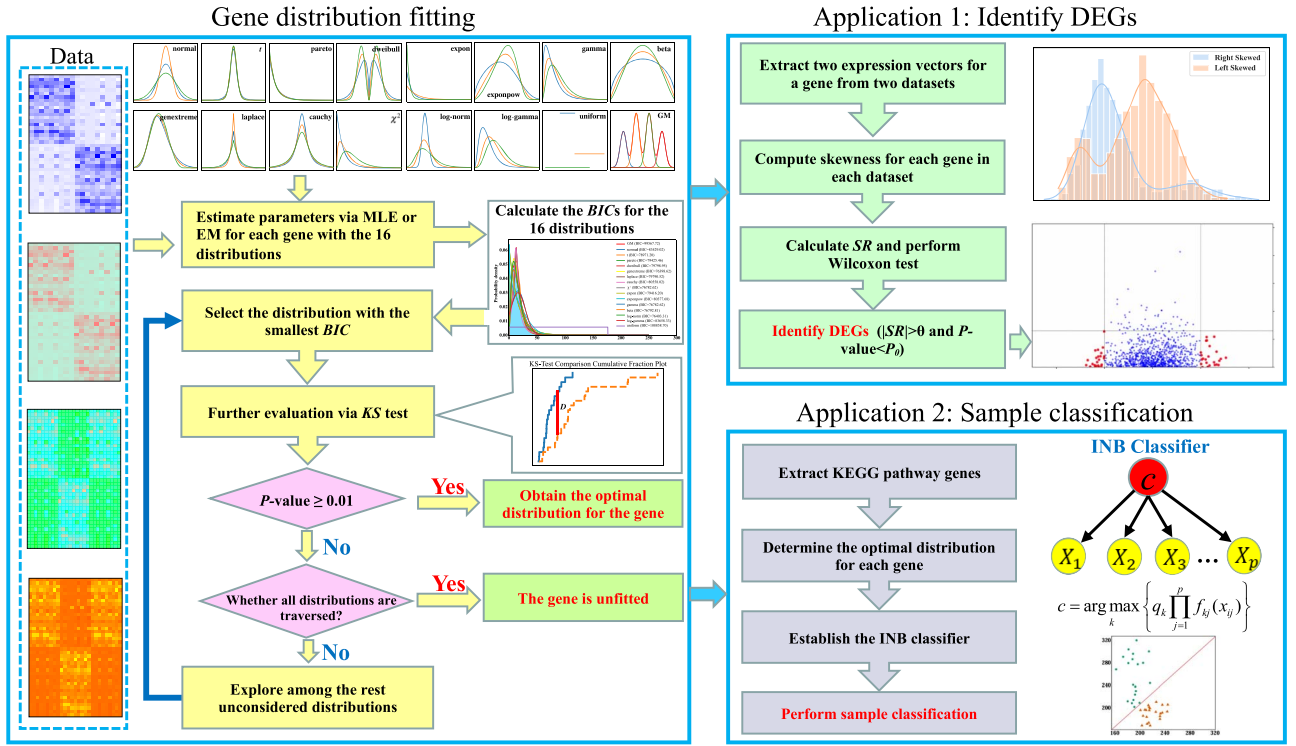


Figure 1. The schematic diagram illustrates the main contents of this paper. Based on the collected omics data from various databases, the dataset is segmented into distinct groups based on clinical information. A total of 16 distributions are evaluated to fit gene distributions across these data groups. A rigorous procedure integrating both the BIC and the KS test is employed for distribution selection. Two key applications are subsequently explored. Firstly, leveraging gene distribution variances between two datasets, a novel metric named SR is introduced to identify Differentially Expressed Genes (DEGs) between these datasets. Secondly, an INB classifier, which integrates gene-specific distributions, is proposed for effective sample classification.

the other distributions are estimated via maximum likelihood estimation (MLE).

Rigorous procedures are developed to determine the optimal distribution for gene j from the 16 distributions (Fig. 1). Initially, the BIC serves as the primary guiding metric:

$$BIC_j = -2 \ln(L_j) + k \ln(n). \quad (2)$$

Here, L_j signifies the maximum value of the likelihood function $\prod_{i=1}^n f(x_{ij})$ for the estimated model (Supplementary material), x_{ij} denotes the expression of the j 'th gene in the i 'th sample; k denotes the number of parameters in the model, and n is the number of samples. Depending on the distribution considered from Table 1, $f(\cdot)$ and k vary accordingly. Secondly, the distribution with the smallest BIC undergoes further scrutiny with the KS test (Supplementary material). If $P \geq 0.01$ from the KS test, it is designated as the optimal distribution for the gene. Otherwise, among the remaining distributions, the one with the smallest BIC is considered next, if $P \geq 0.01$ from the KS test for this subsequent distribution, it is then defined as optimal. If after evaluating all 16 distributions no $P \geq 0.01$ is obtained from the KS test, then the gene is classified as unfitted (Fig. 1).

Skewness ratio: an indicator to measure gene distribution differences across datasets

Suppose that $X_i^{(l)}$ ($l = D_1, D_2$) represents the expression sequences of gene i in dataset l . The skewness of $X_i^{(l)}$ is defined as:

$$Skew(X_i^{(l)}) = \frac{1}{n_i^{(l)}} \sum_{j=1}^{n_i^{(l)}} \left(\frac{x_{ij}^{(l)} - \mu_i^{(l)}}{\sigma_i^{(l)}} \right)^3. \quad (3)$$

Here, $n_i^{(l)}$ represents the number of samples in $X_i^{(l)}$, $\mu_i^{(l)}$, and $\sigma_i^{(l)}$ denote the mean and standard deviation of $X_i^{(l)}$, respectively. Based on skewness, $SR_i^{(D_1, D_2)}$ is defined as follows:

$$SR_i^{(D_1, D_2)} = \frac{Skew(X_i^{(D_1)}) - Skew(X_i^{(D_2)})}{|Skew(X_i^{(D_1)})| + |Skew(X_i^{(D_2)})|}. \quad (4)$$

$SR_i^{(D_1, D_2)} \in [-1, 1]$. $SR_i^{(D_1, D_2)} > 0$ indicates that gene i displays stronger skewness within dataset D_1 , while $SR_i^{(D_1, D_2)} < 0$ represents a more pronounced skewness in dataset D_2 . Additionally, the Wilcoxon test [46] is performed to validate whether the expression profiles of gene i differ significantly between the two datasets. Based on SR and the Wilcoxon test, genes with $|SR| > 0.5$ and P -value < 0.05 are defined as DEGs. The threshold values for $|SR|$ and P can be determined based on data analysis and empirical research, analogous to traditional gene differential expression analysis. Higher $|SR|$ and lower P values result in the selection of fewer genes for subsequent analysis, while lower $|SR|$ and higher P values allow for a broader exploration of gene sets.

The improved naïve Bayes that integrates gene-specific distributions

The NB method is widely employed for sample classification tasks, despite its reliance on several strong assumptions. These assumptions include the independence assumption among different variables and the expectation that all variables adhere to specific distributions [34]. The "naivety" in NB primarily refers to the assumption of conditional independence. Typically, data are assumed to follow normal, multinomial, or Bernoulli distributions when applying the NB method. Despite the challenge of meeting

Table 1. General probability density functions (PDFs) of 16 typical distributions. During parameter fitting, certain distributions are simplified and the data undergo preprocessing to conform to these specific distributions. For further details, refer to the [Supplementary material](#).

Distribution	Notation	PDF	Parameter
Normal	$\mathbf{X} \sim N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	$\sigma > 0, \mu \in \mathbb{R}$
t	$\mathbf{X} \sim t(m)$	$f(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\Gamma\left(\frac{m}{2}\right)} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}$	$m \in \mathbb{Z}^+$
Pareto	$\mathbf{X} \sim \text{Pareto}(\beta, A)$	$f(x) = \frac{\beta A^\beta}{x^{\beta+1}}, x \geq A$	$A > 0, \beta > 0$
Double Weibull	$\mathbf{X} \sim \text{WB}(\lambda, k_1, k_2)$	$f(x) = \frac{k_1 k_2}{\lambda} \left(\frac{x}{\lambda}\right)^{k_1-1} e^{-\left(\frac{x}{\lambda}\right)^{k_1}} \left[1 - e^{-\left(\frac{x}{\lambda}\right)^{k_1}}\right]^{k_2-1}$	$\lambda, k_1, k_2 > 0$
General extr. val.	$\mathbf{X} \sim \text{GEV}(\mu, \sigma, \xi)$	$f(x) = \frac{1}{\sigma} t(x)^{\xi+1} e^{-t(x)}$, $t(x) = \left(1 + \xi \left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}, \xi \neq 0$; $t(x) = e^{-\frac{x-\mu}{\sigma}}, \xi = 0$.	$\mu, \xi \in \mathbb{R}, \sigma > 0$
Laplace	$\mathbf{X} \sim \text{Laplace}(\mu, \lambda)$	$f(x) = \frac{1}{2\lambda} \exp\left(-\frac{ x-\mu }{\lambda}\right)$	$\mu \in \mathbb{R}, \lambda > 0$
Cauchy	$\mathbf{X} \sim C(\lambda, \mu)$	$f(x) = \frac{\lambda}{\pi(\lambda^2 + (x-\mu)^2)}$	$\lambda > 0, \mu \in \mathbb{R}$
χ^2	$\mathbf{X} \sim \chi^2(m)$	$f(x) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{m/2-1} \exp(-\frac{x}{2}), x > 0$	$m \in \mathbb{Z}^+$
Exponential	$\mathbf{X} \sim \text{EXP}(\lambda)$	$f(x) = \lambda e^{-\lambda x}, x > 0$	$\lambda > 0$
Expon. power	$\mathbf{X} \sim \text{EP}(\beta, \alpha, \mu)$	$f(x) = \frac{1}{2\alpha\beta^{1/\beta}\Gamma(1/\beta+1)} \exp\left(-\frac{1}{\beta} \left \frac{x-\mu}{\alpha}\right ^\beta\right)$	$\alpha, \beta > 0, \mu \in \mathbb{R}$
Gamma	$\mathbf{X} \sim \text{Gamma}(\alpha, \beta)$	$f(x) = \frac{\beta^\alpha x^{\alpha-1} \exp(-\beta x)}{\Gamma(\alpha)}, x > 0$	$\alpha, \beta > 0$
Beta	$\mathbf{X} \sim \text{Beta}(\alpha, \beta)$	$f(x) = \frac{\Gamma(\alpha+\beta)x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}, 0 \leq x \leq 1$	$\alpha, \beta > 0$
Log-norm	$\mathbf{X} \sim \text{LN}(\mu, \sigma^2)$	$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\ln x - \mu}{\sigma}\right)^2\right), x > 0$	$\sigma > 0, \mu \in \mathbb{R}$
Log-gamma	$\mathbf{X} \sim \text{LG}(\alpha, \beta)$	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\ln x)^{\alpha-1} \frac{1}{x^{\beta+1}}, x \geq 1$	$\alpha, \beta > 0$
Uniform	$\mathbf{X} \sim U(\alpha, \beta)$	$f(x) = \frac{1}{\beta-\alpha}, \alpha \leq x \leq \beta$	$\alpha, \beta \in \mathbb{R}$
Gaussian mixture	$\mathbf{X} \sim \text{GM}(\mu_i, \sigma_i, w_i, K)$	$f(x) = \sum_{i=1}^K w_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$	$\sigma_i > 0, w_i \geq 0$, $\mu_i \in \mathbb{R}, \sum_{i=1}^K w_i = 1$

these assumptions in practice, NB remains highly competitive in real-world applications [35].

Statistical learning from large-scale omics data reveals that genes often exhibit diverse distribution patterns, with prevalent distributions including log-norm, genextreme, and GM. Thus, incorporating gene-specific distributions is crucial when applying the NB method. Consequently, an Improved Naïve Bayes (INB) classifier is proposed. This classifier first identifies the appropriate distribution for each gene and subsequently integrates these gene-specific distributions into the classification process. The principles underlying the INB can be summarized as follows: given an observation $X_{(i)} = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ for p features, and a class label $c(c = 1, 2, \dots, G)$, based on the NB and conditional probability, the posterior probability classifier can be written as [47]:

$$P(c|X_{(i)}) = \frac{q_c f_c(X_{(i)})}{\sum_{k=1}^G q_k f_k(X_{(i)})}. \quad (5)$$

Here, $\sum_{k=1}^G q_k = 1$; $q_k \in (0, 1)$ is the prior probability of class k , which can be estimated by historical experience or from training data. $f_k(\cdot)$ represents the p dimensional joint PDF for the k 'th class. Since $\sum_{k=1}^G q_k f_k(X_{(i)})$ serves as a normalizing constant, actually, the posterior probability $P(c|X_{(i)}) \propto q_c f_c(X_{(i)})$. Similar to the traditional NB classifier, the p features in $X_{(i)}$ are also assumed to be mutually independent. Thus, $f_c(X_{(i)})$ can be decomposed into the product of marginal PDFs of the p features. That is:

$$f_c(X_{(i)}) = \prod_{j=1}^p f_{cj}(x_{ij}). \quad (6)$$

Here, $f_{cj}(\cdot)$ represents the one-dimensional PDF for the j 'th feature in class c . As a result, the posterior probability from the INB can

be obtained as:

$$P(c|X_{(i)}) = \frac{q_c \prod_{j=1}^p f_{cj}(x_{ij})}{\sum_{k=1}^G q_k \prod_{j=1}^p f_{kj}(x_{ij})}. \quad (7)$$

Sample $X_{(i)}$ can be classified into class c if

$$c = \arg \max_k \left\{ q_k \prod_{j=1}^p f_{kj}(x_{ij}) \right\}. \quad (8)$$

In real-world applications, the prior probability q_k and the PDF $f_{kj}(\cdot)$ can be estimated from the training data. Based on this data, the gene expression distribution for each gene in class k can be estimated, and $f_{kj}(x_{ij})$ can be determined according to the fitted distributions. Given the difficulty in accurately determining cancer incidence rates, the prior probability q_k is typically set equally across different classes k . The detailed computational processes of the INB can be found in the [Supplementary material](#), and the workflow of the INB is shown in Algorithm 1.

It's important to note that the INB requires fitting gene distributions, making its computational complexity higher than that of the traditional NB, especially when dealing with a large number of genes. To mitigate this complexity, only genes involved in KEGG pathways associated with cancers are considered. By leveraging the fitted distributions of pathway genes and observed data from a new sample, the posterior probability that the sample belongs to the associated cancer class can be computed.

Protein-protein interaction network and gene co-expression network

The protein-protein interaction network is curated from the STRING database [48], including only interaction pairs with a confidence score of 0.15 or higher. Gene co-expression relationships are predicted using partial KCC ([Supplementary](#)

Algorithm 1 Workflow of the INB classifier.

- 1: **Input:** Gene expression data $X_{(i)} = (x_{i1}, x_{i2}, \dots, x_{ip})^T$, $i = 1, 2, \dots, n$; class labels c where $c = 1, 2, \dots, G$.
- 2: **Step 1:** Fit gene-specific distributions for each gene j .
- 3: **for** gene $j = 1, 2, \dots, p$ **do**
- 4: Fit class-wise PDF $f_{cj}(x)$.
- 5: **end for**
- 6: **Step 2:** Determine prior probabilities q_k for each class k .
- 7: **Step 3:** Compute joint PDF for each class c according to Eq.(6).
- 8: **for** class $c = 1, 2, \dots, G$ **do**
- 9: $f_c(X_{(i)}) = \prod_{j=1}^p f_{cj}(x_{ij})$.
- 10: **end for**
- 11: **Step 4:** Compute posterior probability for each class c according to Eq.(7).
- 12: **Step 5:** Classify samples according to Eq.(8).
- 13: **Output:** Class label c for sample $X_{(i)}$.

material), with consideration given to gene pairs having an absolute partial KCC greater than 0.1.

Other statistical analysis

The Pearson's χ^2 test [49] is performed when evaluating the independence between rows and columns for a contingency table of the 16 distributions in two datasets. The Wilcoxon rank-sum test [50] is performed to verify whether the expression vectors of a gene in two datasets have significant differences. KEGG and GO enrichment analyses are conducted using the clusterProfiler package [51] in R, employing the hypergeometric test.

Results

Deciphering gene expression distributions using large-scale transcriptomic data

Distributions of cancer driver genes exhibit notable differences between cancerous and normal tissues

Based on the proposed principles for fitting gene distributions (**Materials and Methods**), distributions of genes in cancer and normal tissues are fitted with 16 typical distributions (including normal, t , Pareto, dweibull, genextreme, Laplace, Cauchy, χ^2 , expon, exponpow, gamma, beta, log-norm, log-gamma, uniform, and GM; see Table 1). The contingency table illustrating distributions for the 19,156 genes in the two datasets is presented in Fig. 2A. The ratios of genes in cancer samples that follow the 16 distributions are similar to those in normal samples (KS test, $P = 0.1061$). An independence test between the rows and columns of the contingency table indicates that the distributions of genes in cancer samples are correlated with those in normal samples (χ^2 test, $P < 0.01$). Fitting results for 5,999 (including 3,687 unfitted genes in both datasets) out of the 19,156 genes are consistent across the two datasets. However, 19.32% and 41.81% of genes in normal and cancer samples, respectively, cannot be fitted with any of the 16 distributions. Compared to cancer samples, noticeably more genes in normal samples are fitted with a distribution. Specifically, a considerable number of genes originally following log-norm, GM, beta, χ^2 , gamma, Pareto, normal, exponpow, expon distributions in normal samples are transformed into genextreme, log-norm, GM distributions or cannot be fitted with any of the 16 distributions in cancer samples. This underscores the alteration of gene expression patterns from normal to cancerous states.

Interestingly, the distributions of several known cancer driver genes exhibit significant differences between cancer and normal tissues, such as FAT4, KMT2C, KMT2D, KRAS, PIK3CA, and PTEN [52] (Fig. 2B and **Supplementary Table S1**). It is found that FAT4 follows a Pareto distribution (KS test, $P = 0.0136$) in cancer samples, while it follows exponpow distribution (KS test, $P = 0.8256$) in normal samples. KMT2C exhibits a genextreme distribution (KS test, $P = 0.0696$) and a beta distribution (KS test, $P = 0.2349$) in cancer and normal samples, respectively. Previous studies indicate that mutations in FAT4 may contribute to the development of 24 types of cancers, while mutations in KMT2C may lead to 32 types of cancers [52].

Analysis of unfitted genes reveals that most of them are centered at low expression values and exhibit heavy-tailed distributions influenced by outliers (Fig. 2C, D). The distribution patterns of these unfitted genes indicate that outliers play a significant role in determining whether they can be fitted with one of the 16 distributions. For example, in cancer samples, ATG4A cannot be fitted with any distribution, and the variance of its expression profiles is 87.70. However, after removing outliers (by z-scoring expression levels and replacing absolute scores greater than 3 with the median expression value), its variance reduces to 55.07. Subsequent fitting analysis suggests that a log-normal distribution is optimal (Fig. 2E). This suggests that outliers may indeed contribute to genes being unfitted.

To investigate whether sample sizes affect the optimally fitted distributions, two experiments are conducted. Firstly, 740 samples are randomly selected 20 times from the 10,512 cancer samples. The results show that on average, 20.78% of genes with unknown distributions are observed, which is significantly lower than the proportion observed in all cancer samples ($8010/19156 \approx 41.81\%$). Secondly, an analysis of the evolution of the number of fitted genes indicates a roughly linear decrease with sample sizes (Fig. 2F). These findings suggest that sample sizes may indeed impact the fitting of gene distributions.

Gender-based differences in gene distributions

There are notable gender differences in the incidence and mortality rates of cancer [2, 3, 19, 21–23, 53]. In certain cancers such as bladder, colorectal, and liver cancers, the incidence and mortality rates are higher in males compared to females [2, 3]. These differences may be influenced by hormone levels, gene expression on sex chromosomes, or lifestyle variations among the populations [21–23].

In this section, variations in gene distribution patterns among 4,969 male cancer samples, 5,307 female cancer samples, 370 female normal samples, and 352 male normal samples are explored. Genes in male and female cancer samples primarily exhibit log-normal, genextreme, and GM distributions. Figures 3A and B depict contingency tables of gene distributions for male cancer versus female cancer samples, and male normal versus female normal samples. Independence test between the rows and columns of the contingency tables indicates that the distributions of genes in male samples are correlated with those in female samples (χ^2 test, $P < 2.2e-16$ for both cancer and normal samples). Between the two cancer datasets, only 7,053 genes exhibit identical distributions, while 63.18% of genes show different distributions or are unfitted. Similarly, in male and female normal samples, only 4,755 genes have identical distributions, with 75.18% of genes exhibiting different distributions or being unfitted. The contingency tables for cancer and normal samples differ, with more enriched distributions observed in normal samples. Notably, the numbers of genes following exponpow,

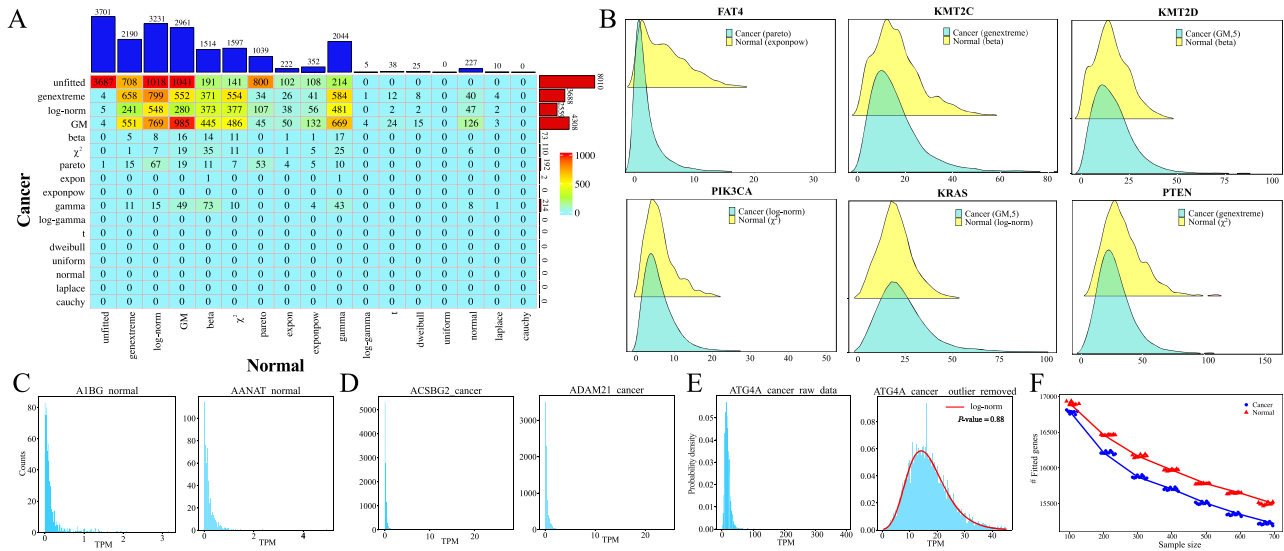


Figure 2. Comparing the fitted distributions between cancer and normal datasets. A. Contingency table summarizing the fitted gene distributions in both datasets. **B.** The fitted distributions of cancer driver genes FAT4, KMT2C, KMT2D, PIK3CA, KRAS, and PTEN in the two datasets. P-values from Wilcoxon test of the fitted distributions from the two datasets for these genes can be seen in **Supplementary Table S1**. Similarly hereinafter. **C.** The histograms of the unfitted genes A1BG and AANAT in normal samples. **D.** The histograms of the unfitted genes ACSBG2 and ADAM21 in cancer samples. **E.** The unfitted gene ATG4A in cancer samples can be fitted with a log-norm distribution after replacing outliers with its median expression values. **F.** The number of fitted genes decreases with increasing sample sizes, considering randomly sampled data from the two datasets under different sample sizes.

gamma, beta, χ^2 , expon, normal, and Laplace distributions are notably higher in normal samples compared to cancer samples.

Interestingly, while similar distribution patterns are observed in both male and female cancer samples, as well as in their respective normal samples (Fig. 3A and B), certain differences are notable. For instance, 839 genes are optimally fitted with normal distributions in female normal samples, whereas only 395 genes exhibit normal distributions in male normal samples. Among the 839 genes following normal distributions in female normal samples, 39, 184, 209, 96, 92, 36, 29, 1, 7, and 5 genes are transformed into GM, exponpow, beta, χ^2 , gamma, genextreme, log-norm, log-gamma, dweibull, and Laplace distributions, respectively, in male normal samples (Fig. 3B). Only 141 genes retain normal distributions in male normal samples. Moreover, several confirmed gender-related carcinogenic genes exhibit completely different distributions in male and female cancer samples, such as PIK3CA, NF1, EIF1AY, IGF1R, NRAS, KDM5D, UTY, and PPP6C (Fig. 3C). In male and female cancer samples, PIK3CA follows a genextreme distribution (KS test, $P = 0.4623$) and a log-norm distribution (KS test, $P = 0.2202$), respectively. Previous studies have reported somatic mutations and amplification of the PIK3CA gene in various human cancers [54–56], including breast cancer, which is particularly prevalent among women. It has also been noted that amplification of the PIK3CA locus 3q25 is more frequent in female kidney renal clear cell carcinoma patients compared to male patients [19]. Another gene, NF1, exhibits a log-norm distribution (KS test, $P = 0.4167$) in male cancer samples and a genextreme distribution (KS test, $P = 0.0470$) in female cancer samples. NF1 encodes neurofibromin 1, a GTPase-activating protein and a key negative regulator of the RAS and PI3K signaling pathways [57]. Moreover, genes KDM5D, UTY, and NF1 have demonstrated gender differences in lung cancer [21], while PPP6C and IGF1R exhibit sex-biased expression in melanoma [22], and the NRAS gene may play sex-specific roles in acute myelogenous leukemia [23].

In summary, analyses of gene distributions in large-scale omics data reveal that PIK3CA, NF1, EIF1AY, IGF1R, NRAS, KDM5D, UTY, and PPP6C display gender-specific expression patterns, indicating their potential crucial roles in gender-related cancers.

Gene distributions exhibit racial disparities

Gene distributions across different racial groups are investigated using a total of 9,340 cancer samples with varying ethnicities, including 7,758 samples from white individuals, 927 samples from black individuals, and 655 samples from Asians. It is observed that in samples from white, black, and Asian individuals, 15.10%, 28.59%, and 29.29% of genes follow log-norm distributions, respectively, while 22.77%, 35.04%, and 20.09% of genes follow genextreme distributions (Fig. 4A). Among the three racial groups, 2,535 genes exhibit identical distributions, whereas the distributions of 4,558 genes are specific to particular racial groups. The most prevalent invariant distributions across the three racial groups are genextreme and log-norm distributions, with a significant number of genes that remained unfitted in the white group also being unfitted in the other two groups (Fig. 4B).

The distribution patterns of several known cancer driver genes, including CTNNB1, ATM, CHEK2, and NCOR1, exhibit racial disparities across datasets (Fig. 4C–F). Gene CTNNB1 follows genextreme distributions in both white and black cancer samples (KS test, $P = 0.1045$ and 0.7404), while it follows a beta distribution in Asian cancer samples (KS test, $P = 0.5109$). Studies on endometrial cancer have shown that CTNNB1 mutations are more frequent in Asian individuals compared to black and white individuals [28]. Gene ATM follows a genextreme distribution in both white and black cancer samples (KS test, $P = 0.6566$ and 0.2219 , respectively), whereas in Asian cancer samples, it follows a log-norm distribution (KS test, $P = 0.7049$). Investigations on bladder cancer have revealed a significantly higher incidence of ATM mutations and associated somatic copy number alterations in white and black populations compared to Asians [26]. Similarly, studies on

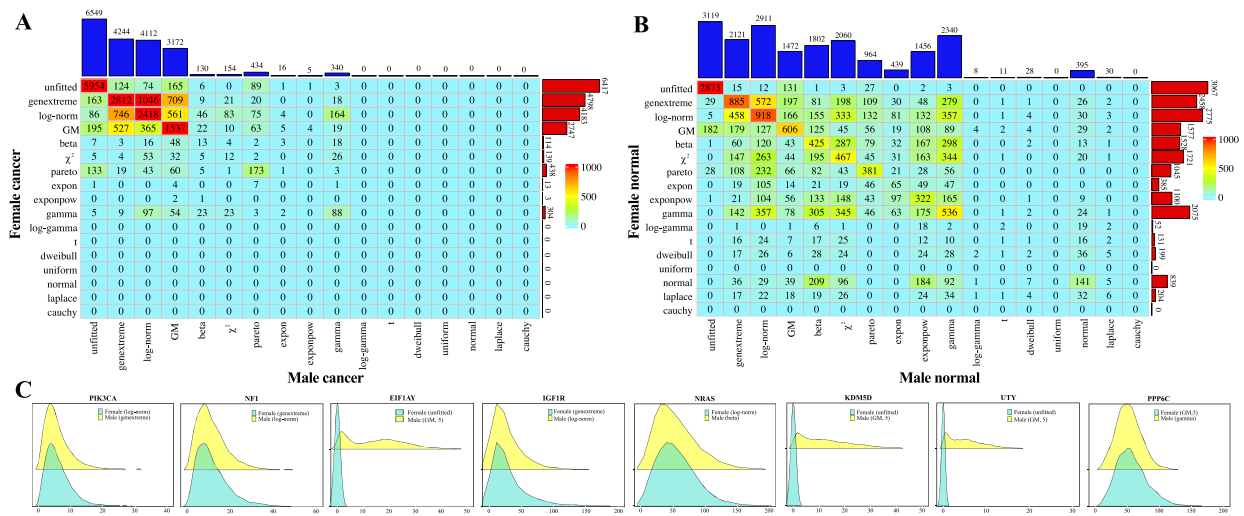


Figure 3. **Disparities in fitted distributions between female and male samples.** **A.** Gene distribution contingency table in female and male cancer samples. **B.** Gene distribution contingency table in female and male normal samples. **C.** The distributions of genes PIK3CA, NF1, EIF1AY, IGF1R, NRAS, KDM5D, UTY, and PPP6C exhibit gender differences in cancer samples between males and females.

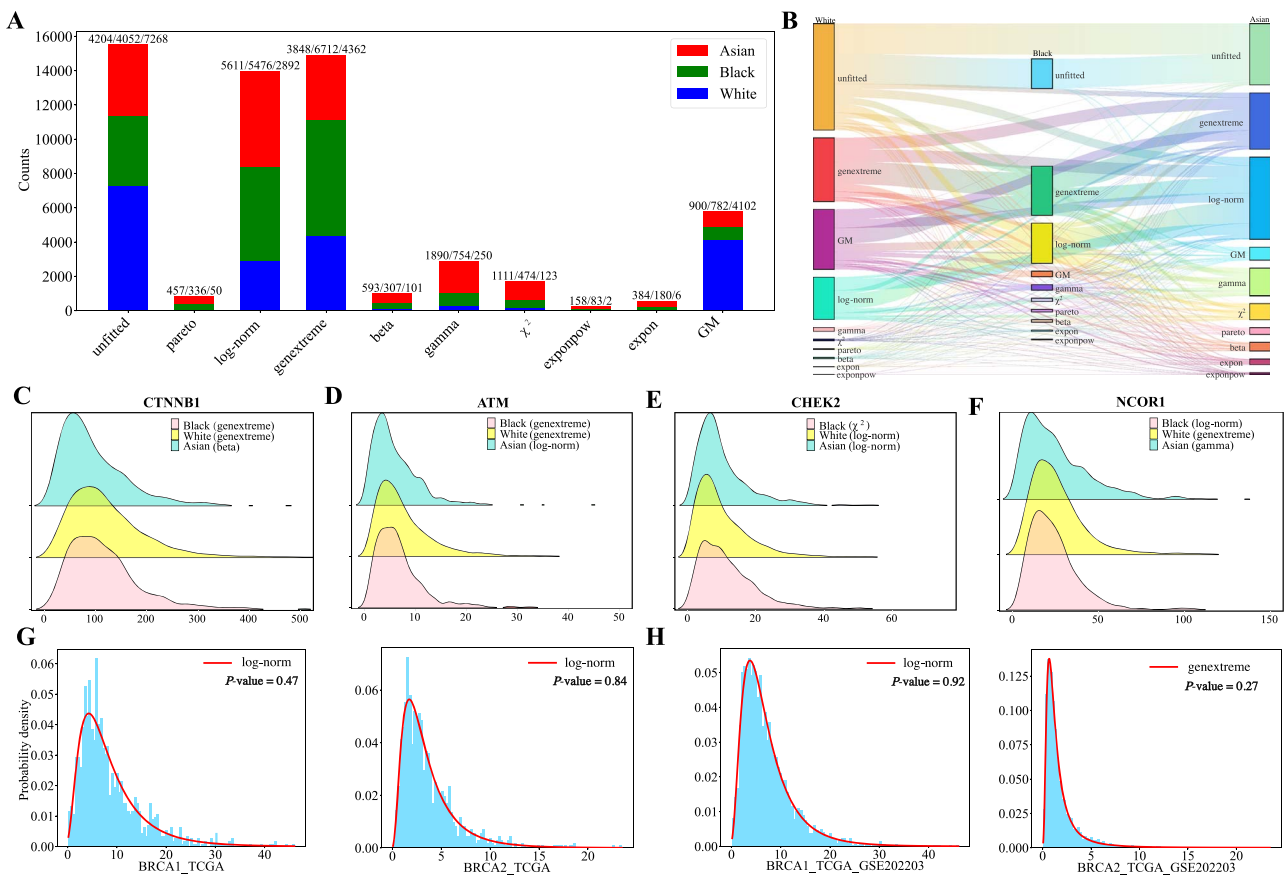


Figure 4. **Racial disparities of gene distributions, and distribution analysis of BRCA1 and BRCA2 in breast cancer datasets.** **A.** The numbers of genes with fitted distributions in black, white and Asian cancer datasets. **B.** Sankey plots illustrating the distributions of genes among the three racial groups. **C-F.** Fitted distributions of genes CTNNB1, ATM, CHEK2, and NCOR1 in the three datasets. **G.** Fitted distributions of BRCA1 and BRCA2 in the breast cancer dataset. **H.** Fitted distributions of BRCA1 and BRCA2 by merging the TCGA data and the GSE202203 breast cancer dataset.

prostate cancer have reported a higher frequency of ATM gene mutations in white and black populations compared to Asians [27]. Gene CHEK2 follows a χ^2 (KS test, $P = 0.8578$), log-norm (KS test, $P = 0.1129$), and another log-norm distribution (KS test, $P = 0.6103$) in black, white, and Asian cancer samples, respectively. Research on breast cancer has indicated that pathogenic variants

of CHEK2 are significantly less frequent in blacks compared to whites [25]. The distribution of the NCOR1 gene in white cancer samples follows a genexreme distribution (KS test, $P = 0.1015$), while in Asian cancer samples it follows a gamma distribution (KS test, $P = 0.8855$), and in black cancer samples it follows a log-normal distribution (KS test, $P = 0.7598$). Previous studies

have reported a significantly higher mutation frequency of NCOR1 in triple-negative breast cancer cases among white populations compared to black populations [24].

In summary, gene distributions reveal racial disparities, with several cancer driver genes exhibiting different distributions across racial groups. These findings suggest that these genes may contribute to variations in cancer incidence among different ethnic groups [58].

Exploring gene distribution patterns across datasets

This subsection further explores two questions: 1) Whether the distributions of cancer driver genes vary with the sources of samples? 2) Whether gene distribution patterns in bulk RNA-seq and scRNA-seq datasets are consistent?

For the first question, a total of 591 canonical cancer driver genes from the NCG database [59] are investigated in both cancer and normal samples. Our analysis reveal that only 120 genes exhibited similar distributions in both datasets, while the remaining 471 genes are fitted with distinct distributions. When comparing distributions between male and female cancer samples, 364 out of the 591 canonical cancer drivers shared the same distribution patterns. This suggests significant distributional differences for a substantial number of canonical cancer driver genes across datasets, which can aid in sample classification. Using the well-known breast cancer suppressor genes BRCA1 and BRCA2 [60, 61] as examples, in 1,118 breast cancer samples, both genes are fitted with log-norm distributions (Fig. 4G). However, when incorporating the bulk RNA-seq dataset GSE202203 with an additional 3,207 breast cancer samples [62], the distribution fitting reveals that BRCA1 still follows a log-normal distribution, while BRCA2 shifts to a genextreme distribution (Fig. 4H). These findings indicate that the sources of samples may influence the fitted distributions of cancer driver genes.

For the second question, two additional scRNA-seq datasets, GSE138536 and GSE202695 [63, 64], are examined. In GSE138536, only 83 out of 10,000 genes can be fitted with one of the 16 distributions, while in GSE202695, only 541 out of 18,273 genes can be fitted with an optimal distribution. The majority of genes in these datasets cannot be fitted with one of the 16 distributions, indicating that the distribution patterns of most genes in scRNA-seq datasets differ from those in bulk RNA-seq datasets.

In conclusion, gene distribution patterns vary significantly across datasets, providing valuable insights for distinguishing between different datasets.

Diversity of gene expression distributions across different developmental stages of cancers

Heterogeneity of gene expression patterns can be influenced by the developmental stages of cancers [65–67]. To explore this heterogeneity, 4,913 out of the 7,758 white cancer samples with developmental stage information are analyzed. The samples are divided into eight subgroups based on gender and cancer stage (Table 2). Each subgroup represents a relatively homogeneous dataset.

Table 2 presents the number of genes fitted with each distribution across the eight datasets. Notably, compared to the results observed in various cancer datasets in the above subsections, a significant increase in the number of genes (approximately 15,000 out of 19,156 genes) exhibiting fitted distributions is observed. These findings suggest that sample heterogeneity may influence gene distribution patterns. Sankey plots illustrating the distributions of genes at different developmental stages

Table 2. The number of genes following each distribution at different developmental stages of cancer for female and male white cancer patients.

Dataset	#samp.	unfit.	t	pareto	log-norm	genextr.	dweibull	beta	log-gamma	gamma	χ^2	lapl.	norm.	exponpow	expon	GM
Fem. stage I	727	3662	107	1093	4526	6541	41	332	0	980	670	22	9	90	175	908
Fem. stage II	829	3794	91	1232	4350	6458	31	383	1	998	577	17	7	60	198	959
Fem. stage III	600	3696	25	1389	4717	5444	3	406	0	1369	920	2	5	143	286	751
Fem. stage IV	212	2953	29	1256	4029	4945	17	582	0	1738	1515	20	62	782	501	727
Male stage I	762	3843	7	1361	4889	5233	0	383	0	1378	782	1	0	71	264	943
Male stage II	675	3874	12	1488	4209	6629	0	396	0	865	544	0	1	107	322	709
Male stage III	679	3874	8	1462	4784	5700	1	386	0	1126	734	0	0	118	316	647
Male stage IV	429	3439	14	1231	4684	6471	4	353	0	1089	704	7	3	196	328	633

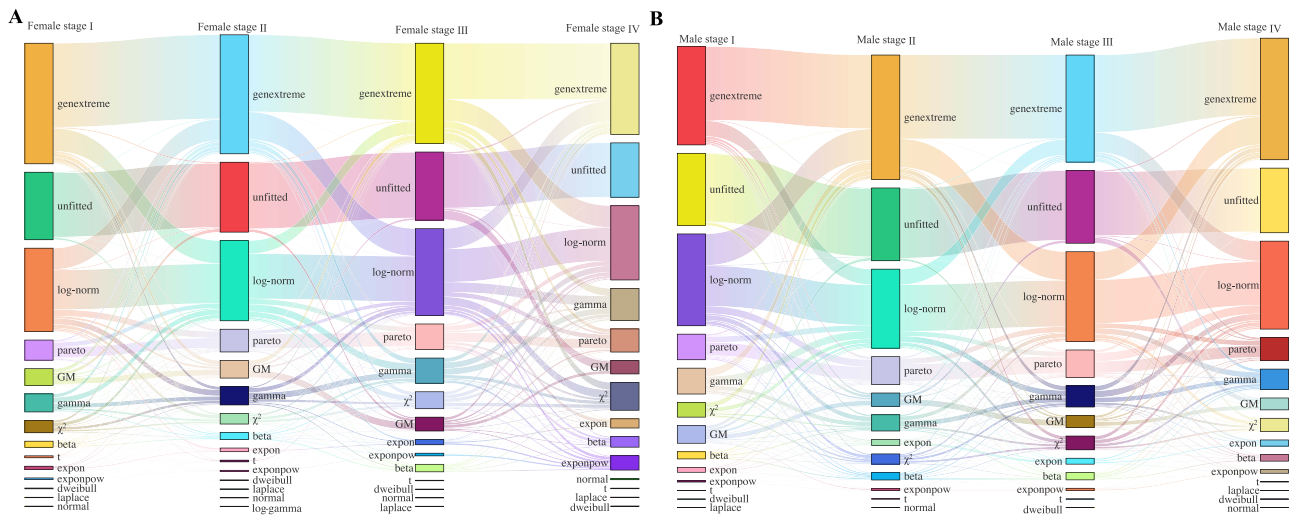


Figure 5. Sankey plots for the distributions of genes at different developmental stages of cancers. **A.** Cases for female cancer samples. **B.** Cases for male cancer samples.

of cancers in female and male cancer samples are depicted in Fig. 5. The figure reveals that many genes follow genextreme and log-norm distributions across different cancer developmental stages, with transitions observed between these distributions at various stages. Comparing across stages, it is observed that in female cancers, a higher proportion of genes in early stages follow genextreme, GM, and dweibull distributions compared to later stages. Conversely, more genes in later stages exhibit gamma, χ^2 , exponpow, and expon distributions. In male cancers, distribution patterns between early and later stages show less obvious distinctions across most distributions. However, both female and male cancers in the early stages tend to have a higher representation of genes following GM distributions compared to the later stages.

Limited occurrence of genes following normal distributions in large-scale transcriptomic data

Comparing the fitted distributions between cancer and normal samples, it is observed that 11,146 genes in cancer samples can be optimally fitted with one of the 16 distributions, predominantly following log-norm, genextreme, and GM distributions. In contrast, for normal samples, 15,455 genes can be optimally fitted with a distribution, mainly encompassing log-norm, GM, genextreme, gamma, χ^2 , beta, and Pareto distributions. Notably, none of the genes in cancer samples are fitted with a normal distribution, while only 227 genes are optimally fitted with normal distributions in normal samples. This raises the question of whether normal distributions are rare in other large-scale omics data.

To investigate whether normal distributions are indeed rare, various datasets are examined (Table 3). Initially, disregarding health status, samples are divided by gender, resulting in 5,321 samples for males and 5,677 samples for females. Distribution fitting shows that 12,617 and 12,767 genes can be optimally fitted with one of the 16 distributions, yet none follows a normal distribution. Next, considering health status, analysis of 4,969 male and 5,307 female cancer samples also reveal no genes following a normal distribution. However, among 352 male and 370 female normal samples, 395 and 839 out of 19,156 genes respectively are optimally fitted with normal distributions. Further examination of racial groups included 655 Asian, 927 Black, and 7,758

White cancer samples, where none of the genes follows a normal distribution. Conversely, among 590 White and 53 Black normal samples, 456 and 454 genes respectively are optimally fitted with normal distributions.

In summary, normal distributions are scarce in large-scale omics data. Across the datasets considered, fewer than 4.5% of genes follow normal distributions in normal samples, with small sample sizes potentially contributing to the appearance of normal distributions (Table 3). Decreased sample sizes reduce the impact of noise and outliers, significantly reducing the number of unfitted genes, which may allow for a few genes to exhibit normal distributions.

Identifying informative genes via skewness of gene distributions

Identifying DEGs distinguishing cancer and normal samples using skewness ratio

Recognizing skewed gene distributions in cancer datasets and variations across different datasets, a statistical measure called the Skewness Ratio (SR) based on skewness [68] and the Wilcoxon rank-sum test [50] is proposed (Materials and Methods). The SR facilitates assessment of skewness variations in gene expression profiles between datasets. Based on the proposed SR metric, and the expression profiles of the 19,156 genes across 10,152 cancer samples and 730 normal samples, 16,492 genes show significant differential expression with $|SR| > 0.5$ and $P < 0.05$ (Wilcoxon rank-sum test). Notably, well-known oncogenes such as FAT4, KMT2C, KRAS, and PIK3CA display high $|SR|$ values of 0.7062, 0.5315, 0.9152, and 0.7746, respectively, indicating highly skewed distributions. To validate the biological relevance of genes with high $|SR|$ and to compare their identification with traditional methods like DESeq2 and edgeR [41, 42], the top 500 genes identified by the three methods are examined. Among these, 469 genes are uniquely identified by SR, and notably, 17 of them are confirmed as canonical cancer drivers [59]. KEGG enrichment analysis reveals that these 469 genes are significantly enriched in cancer-related pathways, including Ras signaling, PI3K-Akt signaling, and Renal cell carcinoma pathways (Fig. 6A, Supplementary Material and Fig. S1).

Based on the STRING database [48] and using the partial Kendall correlation coefficient (KCC) [69], a network is constructed

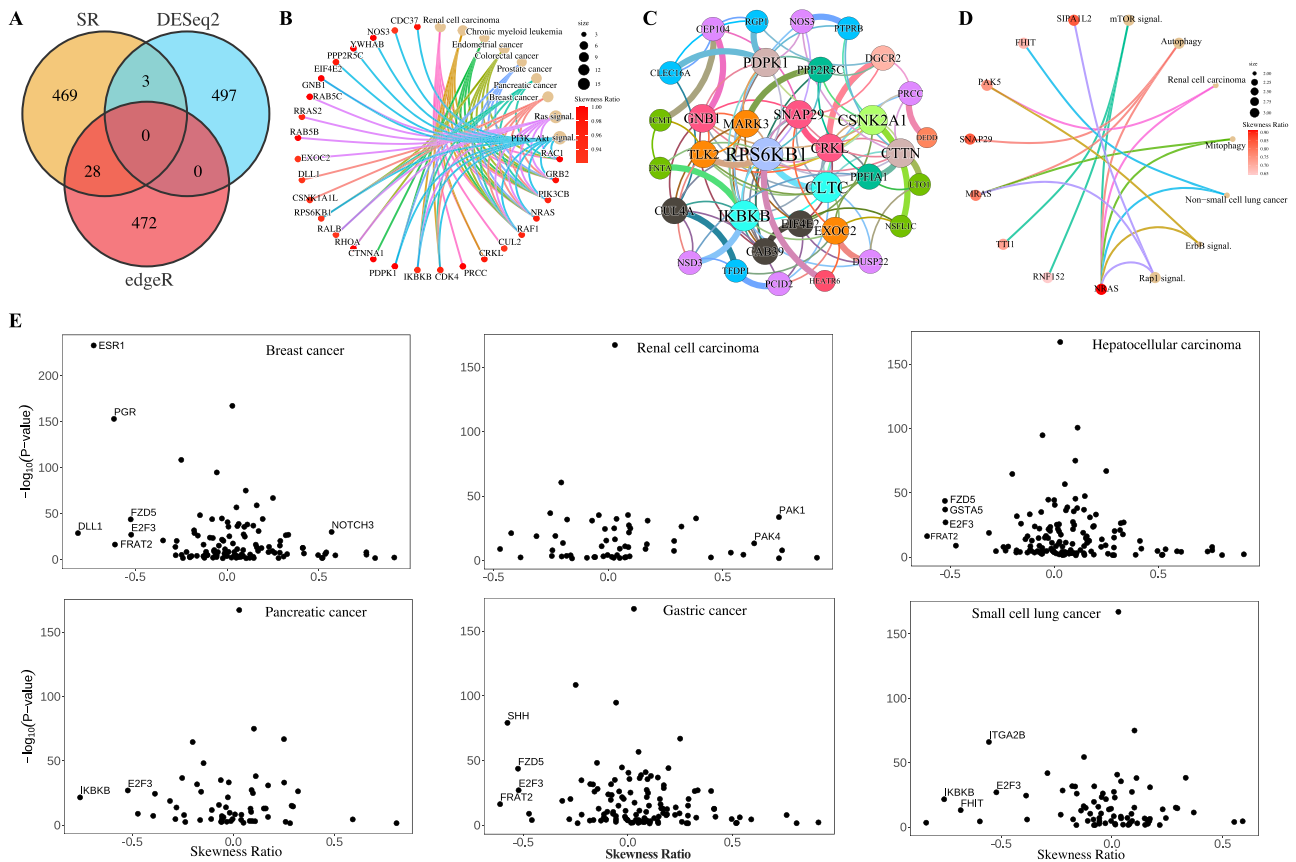


Figure 6. Performances of the SR in identifying DEGs. **A.** Venn diagram for DEGs screened by SR, DESeq2, edgeR when comparing between cancer and normal samples. **B.** KEGG enrichment analysis for the top 469 genes (Not identified in the top 500 by DESeq2 and edgeR) that are uniquely identified by SR when comparing between cancer and normal samples. Significantly enriched pathways ($P < 0.05$) related to cancers are displayed. Similarly hereinafter. **C.** Network analysis of the 469 genes uniquely selected by |SR|. Thin links indicate interactions curated from the STRING database, while thick links represent connections also predicted by the partial KCC (absolute partial correlation greater than 0.1). **D.** KEGG enrichment analysis for the top 93 genes that are uniquely identified by SR when comparing between male and female cancer samples. **E.** The SR versus P values from the Wilcoxon test for genes that involve in the breast cancer, renal cell carcinoma, hepatocellular carcinoma, pancreatic cancer, gastric cancer, and small cell lung cancer pathways. Here, only genes with $P < 0.05$ are considered, and names of genes with $-\log_{10}(P) > 10$ and $|SR| > 0.5$ are displayed.

for the 469 genes uniquely identified via |SR| (Fig. 6B, Supplementary material and Fig. S2). Figure 6B highlights several gene pairs connected by protein-protein interactions and predicted by partial correlations. Notably, interactions include MARK3 and PPP2R5C ($KCC = 0.1468$), FNTA and IKBKB ($KCC = 0.2028$), NSD3 and IKBKB ($KCC = 0.1154$), HEATR6 and RPS6KB1 ($KCC = 0.1333$), CLTC and RPS6KB1 ($KCC = 0.1158$), DUSP22 and EXOC2 ($KCC = 0.1209$), SNAP29 and CRKL ($KCC = 0.1409$), CRKL and DGCR2 ($KCC = 0.1058$), E1F4E2 and CAB39 ($KCC = 0.1032$), GNB1 and CEP104 ($KCC = 0.1922$), PTPRB and NOS3 ($KCC = 0.1950$), RGP1 and PDPK1 ($KCC = 0.1072$), CLEC16A and PDPK1 ($KCC = 0.1020$), DEDD and PRCC ($KCC = 0.1760$). Figure 6A illustrates that PPP2R5C, IKBKB, RPS6KB1, EXOC2, CRKL, EIF4E2, GNB1, NOS3, PDPK1, and PRCC are involved in multiple cancer pathways, suggesting that MARK3, FNTA, HEATR6, CLTC, DUSP22, SNAP29, DGCR2, CAB39, CEP104, NSD3, PTPRB, RGP1, CLEC16A, DEDD, and TLK2 may also play significant roles in these cancers.

Identifying DEGs distinguishing between female and male cancer patients using skewness ratio

Comparing the 5,307 female cancer samples with the 4,969 male cancer samples, 15,406 genes that exhibit significant skewness differences ($|SR| > 0.5$ and $P < 0.05$) are identified. As discussed in the former subsection, genes EIF1AY, IGF1R, KDM5D, NRAS,

PPP6C, and UTY have been previously reported to demonstrate gender differences in lung cancer, gastric cancer, liver cancer, melanoma, and acute myeloid leukemia. The skewness ratio (SR) values for these genes are 0.8923, 0.7595, 0.8302, 0.9112, 0.8111, and 0.8663, respectively, indicating substantial variations in skewness between the female and male cancer datasets.

The top 500 DEGs identified by DESeq2, edgeR, and SR are further analyzed. Among these, SR uniquely identified 93 genes, which show significant enrichment in several cancer pathways such as non-small cell lung cancer and renal cell carcinoma (Fig. 6C, Supplementary material and Fig. S3). Previous studies have indicated gender differences in these types of cancers [70, 71].

Among the 15,406 genes exhibiting significant skewness differences between female and male cancer samples, those involved in breast cancer, gastric cancer, hepatocellular carcinoma, pancreatic cancer, renal cell carcinoma, and small cell lung cancer pathways are specifically examined. Scatter plots depicting SR versus $-\log_{10}(P)$ for these pathway-associated genes are presented in Fig. 6D-I. Our findings illustrate that each pathway includes several genes displaying notable skewness differences between female and male patients, suggesting these genes could potentially serve as important markers for gender prediction in samples.

Table 3. The number of genes that follow normal distributions in various datasets.

Dataset	Norm.	Can.	Male	Fem.	Male nor.	Male can.	Fem. nor.	Fem. can.	Asian can.	Black can.	Black nor.	White can.	White nor.
Samp. size	740	10512	5321	5677	352	4969	370	5307	655	927	53	7758	590
Fit. genes	15455	11146	12617	12767	16037	12607	16089	12739	14952	15104	13556	11888	13559
Norm. distr.	227	0	0	0	395	0	839	0	0	0	454	0	456
Prop.	1.19%	0	0	0	2.06%	0	4.38%	0	0	0	2.37%	0	2.38%

Improved naïve Bayes integrating gene-specific distributions for sample classification

Performance of the improved naïve Bayes in simulated data

By integrating gene-specific distributions into the traditional NB framework [72], we propose an improved NB (INB) classifier for sample classification (**Materials and Methods**). To validate the effectiveness of the INB model in simulated data, we conduct two simulations using randomly generated datasets with specific distributions.

Firstly, we explored the computational processes of the INB model through two simple examples, where only two variables with specific distributions are considered (detailed in the [Supplementary material](#)). These examples revealed that the traditional NB model can suffer from compromised classification accuracy when incorrect distribution assumptions are made. Conversely, the INB model demonstrated superior performance in handling categorical data with diverse distributions.

Secondly, two datasets, each containing 400 samples and 200 variables, are generated to represent two distinct sample types. Within each dataset, 100 variables follow a generalized extreme value distribution, while the remaining 100 variables follow a log-normal distribution. These datasets served as training sets for the model. Subsequently, 80 samples are generated as test sets. To compare the performance of the INB and the traditional NB, the actual distribution for each variable is utilized in the INB model, while the NB model assumes all variables followed normal distributions. The F1-score (F1) and Youden index (YI) for each method are averaged over 50 random simulation runs. Results indicate that the INB achieves superior classification performance, with an average F1 of 0.9774 and a YI of 0.9548, compared to the NB's average F1 of 0.8536 and YI of 0.71. This clearly demonstrates the benefit of incorporating gene-specific distributions in the INB model. Notably, the experiment also reveals that even when datasets do not conform to normal distributions, the NB model still exhibits reasonable classification performance.

Performance of the improved naïve Bayes in sample classification

To assess the classification performance of the INB in real-world datasets, datasets for hepatocellular carcinoma (HCC), breast cancer (BRCA), kidney cancer (KICA), lung adenocarcinoma with lung squamous cell carcinoma (LUAD_LUSC), and thyroid carcinoma with bladder urothelial carcinoma (THCA_BLCA) are considered (**Materials and Methods**). These datasets include specific cancer samples alongside their corresponding normal samples, as well as datasets with mixed cancer samples. For datasets with specific cancer samples, features in the classifiers are derived from KEGG pathway genes associated with the respective cancers. In contrast, for datasets with mixed cancer samples (LUAD_LUSC and THCA_BLCA), the feature set includes the union of KEGG pathway genes corresponding to both cancer subtypes. It's important to note that in the HCC, BRCA, and KICA datasets, the cancer samples represent a subset of specific cancer cases, with normal samples typically sourced from healthy tissues of corresponding cancer patients. Conversely, in the LUAD_LUSC and THCA_BLCA datasets, samples are sourced from different patients.

To validate the advantages of using KEGG pathway genes and to assess the classification performance of the INB using gene-specific distributions, two control experiments are conducted: 1) INBFD: The INB using only five types of distributions (expon, pareto, log-gamma, uniform, and normal distributions). 2) INBRSG: Randomly selecting an equivalent number of genes from

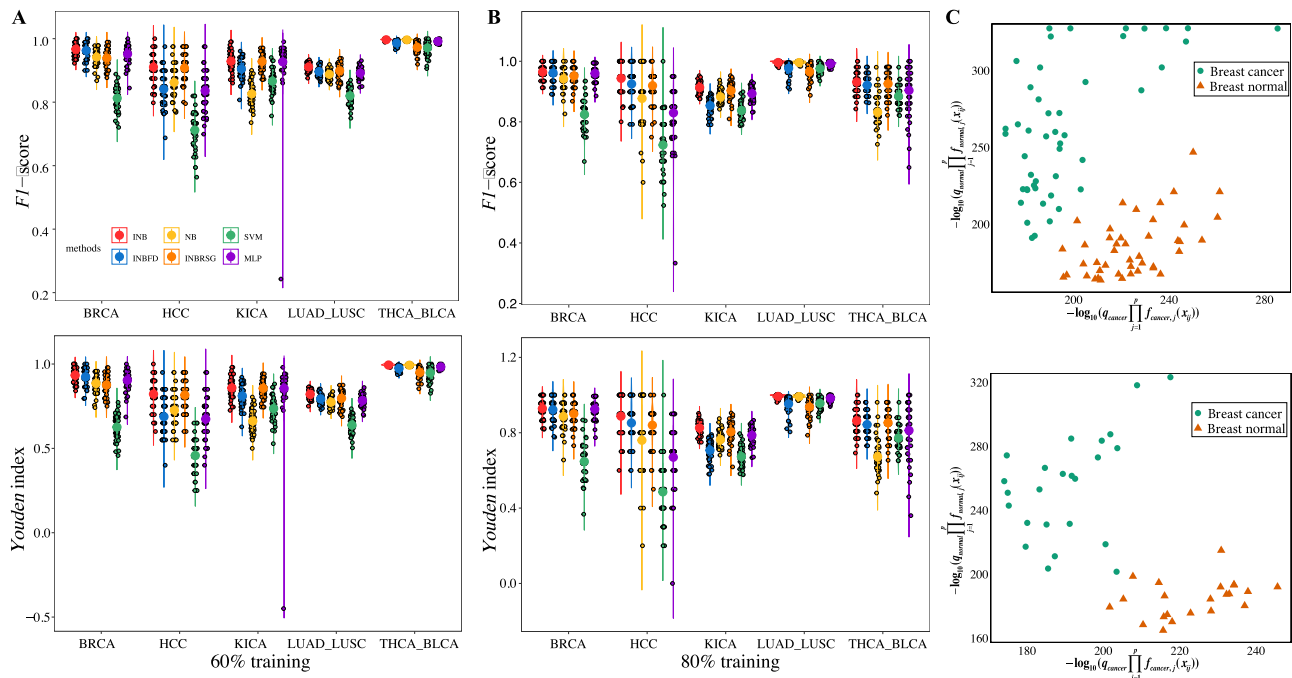


Figure 7. **Performances of the INB classifier in sample classification.** **A.** F1 and YI for the INB, NB, INBFD, INBRSG, SVM, and MLP using 60% training data across the five datasets. **B.** F1 and YI for the INB, NB, INBFD, INBRSG, SVM, and MLP using 80% training data across the five datasets. **C.** Sample classification using INB in the breast cancer dataset with 60% training data and 80% training data. Here, $q_k \prod_{j=1}^p f_{kj}(x_{ij})$ is transformed using $-\log_{10}$.

the dataset for sample classification. Additionally, the Support Vector Machine (SVM) and Multilayer Perceptron (MLP) classifiers are included for comparison [73, 74]. For each experiment, 80% and 60% of randomly extracted samples are used as training datasets, while the remaining samples serve as the testing dataset.

The average F1 and YI from 50 simulation runs indicate that the INB outperforms the NB across the five datasets. In all cases, the F1 score of the INB exceeds 0.91 and the YI surpasses 0.82 (see Table 4, Fig. 7A and B). Notably, in the KICA dataset with 80% training data, the INB achieves an F1 score nearly 12% higher than that of the NB, and the YI is 28% higher. Scatter plots of $-\log_{10}(q_k \prod_{j=1}^p f_{kj}(x_{ij}))$ for breast cancer and normal datasets illustrate that the INB effectively distinguishes between cancer and normal samples. Comparisons among INB, NB, INBFD, INBRSG, SVM, and MLP (Table 4 and Fig. 7C) consistently show that the INB maintains superior performance in sample classification, underscoring the benefit of using accurate gene-specific distributions. In summary, the proposed INB, which incorporates optimal gene-specific distributions, demonstrates superior classification performance compared to the NB, SVM, and MLP.

Performance of the improved Naïve Bayes in other testing data

Based on the previously trained INB classifiers from the breast cancer and kidney cancer datasets in the aforementioned section, some extra data for the two cancers are explored. The extra data includes the previously unused samples from the TCGA for the two cancer, as well as the breast cancer dataset GSE202203 from NCBI [62]. The unused TCGA breast cancer data consist of 1,002 cancer samples and 3 normal samples. The unused TCGA kidney cancer data include 773 cancer samples and 3 normal samples. The GSE202203 dataset contains 3,207 samples for breast cancer.

For the extra TCGA breast cancer samples, the INB achieves an F1 score of 0.9882 and a YI of 0.9810, while the NB method

achieves an F1 score of 0.9574 and a YI of 0.9232. In the breast cancer dataset GSE202203 from NCBI, the INB achieves an F1 score of 0.9992, surpassing the NB's F1 score of 0.9782. Regarding the extra TCGA kidney cancer samples, the INB achieves an F1 score of 0.9955 and a YI of 0.6628, compared to the NB's F1 score of 0.8275 and a YI of 0.7102. Although the YI of the INB is slightly lower than that of the NB for the extra TCGA kidney cancer samples, the F1 score of the INB is significantly higher, underscoring the practical efficacy of the INB. Despite relying solely on the fitted distributions of genes within associated KEGG pathways, the INB notably enhances classification accuracy.

Discussion

Cancer remains a formidable global health challenge with significant mortality rates. Current research often assumes that genes conform to singular distributions such as normal, multinomial, or Bernoulli distributions. However, the true distributions of genes in large-scale omics data, including their variations across genders and racial groups, are not well understood. Advances in high-throughput sequencing have facilitated the exploration of gene expression distributions across different datasets derived from TCGA, categorized by cancer types, racial demographics, and gender. Rigorous fitting criteria, including BIC and KS tests, reveal distinct distribution patterns among known cancer driver genes, gender-specific genes, and racial-specific genes. Notably, a majority of genes exhibit highly skewed distributions, with cancer samples preferring genextreme, log-norm, and GM distributions, while normal samples show preferences for gamma, beta, χ^2 , Pareto, exponpow, and expon distributions. Importantly, over 95.5% of genes deviate from normal distributions, a trend particularly pronounced in larger cancer datasets.

Investigations into gene expression heterogeneity show that cancer samples exhibit more asymmetric distributions compared to their normal counterparts. For instance, while only 275 genes

Table 4. Mean values and standard deviations of F1 and YI with 80% and 60% training data in the five datasets based on 50 random simulation runs.

Dataset	80% training					60% training						
	INB	NB	INBFD	INBRSG	SVM	MLP	INB	NB	INBFD	INBRSG	SVM	MLP
F1	BRCA	0.9647 (0.0245)	0.9421 (0.0340)	0.9617 (0.0219)	0.9515 (0.0278)	0.8246 (0.0479)	0.9670 (0.0196)	0.9417 (0.0249)	0.9625 (0.0225)	0.9392 (0.0284)	0.8126 (0.0367)	0.9518 (0.0284)
	HCC	0.9437 (0.0515)	0.8773 (0.0975)	0.9252 (0.0586)	0.9195 (0.0532)	0.7235 (0.0901)	0.9106 (0.0436)	0.8619 (0.0494)	0.8410 (0.0551)	0.9075 (0.0438)	0.7121 (0.0556)	0.8358 (0.0603)
	KICA	0.9314 (0.0336)	0.8317 (0.0474)	0.9211 (0.0291)	0.9258 (0.0399)	0.8847 (0.0318)	0.9287 (0.0294)	0.8263 (0.0344)	0.9047 (0.0304)	0.9278 (0.0265)	0.8661 (0.0351)	0.9269 (0.1020)
	LUAD_LIUSC	0.9133 (0.0174)	0.8822 (0.0218)	0.8536 (0.0261)	0.9029 (0.0261)	0.8376 (0.0258)	0.9107 (0.0137)	0.8877 (0.0143)	0.8965 (0.0144)	0.8990 (0.0230)	0.8200 (0.0259)	0.8917 (0.0171)
	THCA_BLCA	0.9959 (0.0049)	0.9958 (0.0043)	0.9772 (0.0213)	0.9655 (0.0251)	0.9758 (0.0144)	0.9967 (0.0029)	0.9963 (0.0028)	0.9860 (0.0088)	0.9727 (0.0158)	0.9712 (0.0214)	0.9915 (0.0047)
YI	BRCA	0.9286 (0.0504)	0.8864 (0.0668)	0.9225 (0.0589)	0.9006 (0.0571)	0.6474 (0.0916)	0.9335 (0.0397)	0.8850 (0.0494)	0.9239 (0.0460)	0.8762 (0.0579)	0.6261 (0.0694)	0.9039 (0.0567)
	HCC	0.8880 (0.1023)	0.7600 (0.1895)	0.8520 (0.1147)	0.8400 (0.1050)	0.486 (0.1539)	0.8220 (0.0864)	0.7260 (0.0970)	0.6880 (0.1052)	0.8160 (0.0866)	0.458 (0.0966)	0.674 (0.1192)
	KICA	0.8638 (0.0660)	0.6728 (0.0885)	0.8431 (0.0575)	0.8531 (0.0780)	0.7722 (0.0618)	0.8584 (0.0577)	0.6620 (0.0637)	0.8114 (0.0590)	0.8561 (0.0525)	0.7358 (0.0678)	0.8553 (0.1950)
	LUAD_LIUSC	0.8256 (0.0350)	0.7633 (0.0438)	0.7067 (0.0519)	0.8049 (0.0525)	0.6731 (0.0506)	0.8207 (0.0277)	0.7738 (0.0288)	0.7934 (0.0283)	0.7970 (0.0458)	0.6379 (0.0499)	0.7836 (0.0337)
	THCA_BLCA	0.9926 (0.0089)	0.9921 (0.0079)	0.9532 (0.0445)	0.9568 (0.0462)	0.9564 (0.0260)	0.9940 (0.0052)	0.9930 (0.0053)	0.9742 (0.0164)	0.9500 (0.0288)	0.9474 (0.0388)	0.9827 (0.0099)

in normal samples display symmetric distributions (227 normal distributions, 38 t distributions, and 10 Laplace distributions), this pattern is absent in all cancer samples, including female and male subsets. These findings underscore how increased heterogeneity in gene expression profiles contributes to distinct distribution patterns in cancer versus normal samples.

Further exploration includes comparing gene distributions between bulk and scRNA-seq datasets. Notably, the majority of genes in scRNA-seq datasets cannot be fitted to any of the 16 distributions considered. This discrepancy is attributed to the “dropout” phenomenon in scRNA-seq, where a gene may be expressed in some cells but undetected in others due to low RNA input and stochastic gene expression.

Motivated by these findings on gene distribution diversity, a new metric called SR is introduced to measure skewness variation across datasets. This metric evaluates gene importance in distinguishing between datasets based on their distributional deviations. Additionally, the Wilcoxon test verifies significant expression differences for the considered genes between datasets. The SR provides a holistic view of gene significance, contrasting with traditional methods like $\log_2(FC)$, which only assess average expression differences and are susceptible to outlier impacts. Numerical simulations across cancer and normal datasets demonstrate that top-ranked genes by SR frequently include canonical cancer drivers and are enriched in cancer-related pathways. Furthermore, in gender-specific cancer analyses, SR effectively identifies DEGs associated with gender-biased cancers.

To harness the practical utility of gene-specific distributions, an INB classifier is developed. Unlike traditional NB, which assumes genes follow standard distributions, INB fits optimal distributions for each gene across datasets, enhancing classification accuracy. The methodological approach of leveraging variable-specific distributions can be extended to other algorithms to bolster their classification capabilities. Future research could also explore additional distributions beyond the 16 distributions, such as the skew generalized normal distribution [75]. Additionally, developing methods that relax the independence assumption in the NB and INB models [76, 77] presents a promising avenue for future work. Our method involves fitting multiple distribution models, while applying MLE and the EM algorithm to optimize the parameters of these distributions. These computational processes, especially when dealing with large-scale gene expression data, may require considerable computational resources. For INB, running a simulated dataset with 20 samples and 10,000 variables on an Intel(R) Core (TM) i7-8700 CPU @ 3.20GHz 3.19 GHz takes 8,583 seconds. Although the computational time is manageable for ordinary computers, to further improve computational efficiency, several optimization strategies can be considered, such as vectorized operations, parallel computing, and a distribution pre-screening strategy. Additionally, although our method primarily relies on CPU computation, for larger datasets, it may be beneficial to migrate some computational tasks to GPU, particularly in terms of parallelization and matrix operations, where GPU computing can significantly enhance efficiency.

Challenges and avenues for future investigation include exploring the performance of SR when using fitted gene distributions, addressing the computational complexity of INB for large gene sets by focusing on KEGG pathway genes, and integrating gene network information to refine classification algorithms. Beyond differential expression and sample classification, future research could delve into inferring gene regulatory networks [78]. Moreover,

expanding this study to encompass multi-omics data promises further insights, which is a direction to be pursued in future work.

Conclusions

This paper deciphers gene distributions in large-scale RNA-seq data and explores their applications. Extensive statistical analysis reveals that genes in normal samples exhibit more diverse distributions compared to those in cancer samples. Genes in cancer samples tend to prefer distributions such as genextreme, log-norm, and GM, whereas genes in normal samples favor gamma, beta, χ^2 , Pareto, exponpow, and expon distributions. Remarkably, less than 4.5% of genes in normal samples follow normal distributions. Some well-known cancer drivers and genes with gender and racial differences exhibit distinct distribution characteristics across datasets. Based on the differences in gene skewness between two datasets, a metric called SR is introduced, which provides an alternative method to explore DEGs. Numerical simulations across multiple datasets demonstrate that the SR measure offers advantages over traditional methods like DESeq2 and edgeR. Furthermore, an INB classifier is developed, which incorporates gene-specific distributions and shows higher classification accuracy and stability compared to the traditional NB classifier in various datasets. These investigations not only deepen our understanding of gene distribution patterns but also introduce novel approaches for DEG screening and sample classification.

Key Points

- Gene expression patterns in large-scale transcriptomic data are deciphered.
- A novel skewness-based metric is proposed to screen differentially expressed genes across two datasets.
- An improved naïve Bayes method incorporating gene-specific distributions can well realize sample classification.

Acknowledgments

The datasets used in this paper are in part based upon the data generated by the TCGA Research Network (<https://www.cancer.gov/tcga>) and the GEO database (<https://www.ncbi.nlm.nih.gov/geo>). This work is supported by the National Natural Science Foundation of China under Grant 61773153 and by the Natural Science Foundation of Henan Province under Grant 202300410045.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Competing interests

No competing interest is declared.

Author contributions statement

Conceptualization: W.P.; Methodology: W.P., C.S.J., G.H.P.; Investigation: C.S.J., W.P.; Visualization: C.S.J., W.P., G.H.P., Z.Y.J.; Supervision: W.P.; Writing—original draft: C.S.J., W.P.; Writing—review & editing: W.P., C.S.J.; Funding acquisition: W.P.

References

- Deng Q, Ramsköld D, Reinius B. et al. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 2014;**343**:193–6. <https://doi.org/10.1126/science.1245316>.
- Dong M, Cioffi G, Wang J. et al. Sex differences in cancer incidence and survival: a pan-cancer analysis. *Cancer Epidemiol Biomarkers Prev* 2020;**29**:1389–97. <https://doi.org/10.1158/1055-9965.EPI-20-0036>.
- Naugler WE, Sakurai T, Kim S. et al. Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production. *Science* 2007;**317**:121–4. <https://doi.org/10.1126/science.1140485>.
- Scharl T, Grün B, Leisch F. Mixtures of regression models for time course gene expression data: evaluation of initialization and random effects. *Bioinformatics* 2010;**26**:370–7. <https://doi.org/10.1093/bioinformatics/btp686>.
- Lai Y, Zhang F, Nayak TK. et al. An efficient concordant integrative analysis of multiple large-scale two-sample expression data sets. *Bioinformatics* 2017;**33**:3852–60. <https://doi.org/10.1093/bioinformatics/btx061>.
- Cai G, Zhao W, Zhou Z. et al. MATTE: a pipeline of transcriptome module alignment for anti-noise phenotype-gene-related analysis. *Brief Bioinform* 2023;**24**:bbad207. <https://doi.org/10.1093/bib/bbad207>.
- Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;**11**:31–46. <https://doi.org/10.1038/nrg2626>.
- Ritchie MD, Holzinger ER, Li R. et al. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* 2015;**16**:85–97. <https://doi.org/10.1038/nrg3868>.
- De Torrenté L, Zimmerman S, Suzuki M. et al. The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. *BMC Bioinform* 2020;**21**:1–18.
- Chen S, Yang S, Wang P. et al. Two-stage penalized algorithms via integrating prior information improve gene selection from omics data. *Phys A Stat Mech Appl* 2023;**628**:129164. <https://doi.org/10.1016/j.physa.2023.129164>.
- Lyu J, Li JJ, Jianzhong S. et al. DORGE: discovery of oncogenes and tumor suppressor genes using genetic and epigenetic features. *Sci Adv* 2020;**6**:eaba6784.
- Kossinna P, Cai W, Xuewen L. et al. Stabilized core gene and pathway election uncovers pan-cancer shared pathways and a cancer-specific driver. *Sci Adv* 2022;**8**:eabo2846. <https://doi.org/10.1126/sciadv.abo2846>.
- Cho J, Baik B, Nguyen HCT. et al. Characterizing efficient feature selection for single-cell expression analysis. *Brief Bioinform* 2024;**25**:bbae317. <https://doi.org/10.1093/bib/bbae317>.
- Meng Q, Wei L, Ma K. et al. scDecouple: decoupling cellular response from infected proportion bias in scCRISPR-seq. *Brief Bioinform* 2024;**25**:bbae011. <https://doi.org/10.1093/bib/bbae011>.
- Yang S, Chen S, Wang P. et al. TSPLASSO: a two-stage prior LASSO algorithm for gene selection using omics data. *IEEE J Biomed Health Inform* 2024;**28**:526–37. <https://doi.org/10.1109/JBHI.2023.3326485>.
- Wang P, Chen S, Yang S. Recent advances on penalized regression models for biological data. *Mathematics* 2022;**10**:3695. <https://doi.org/10.3390/math10193695>.
- Amundadottir LT, Thorvaldsson S, Gudbjartsson DF. et al. Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family. *PLoS Med* 2004;**1**:e65. <https://doi.org/10.1371/journal.pmed.0010065>.
- Yang J, Hainaut P, Gores GJ. et al. A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nat Rev Gastroenterol Hepatol* 2019;**16**:589–604. <https://doi.org/10.1038/s41575-019-0186-y>.
- Janku F, Wheler JJ, Westin SN. et al. PI3K/AKT/mTOR inhibitors in patients with breast and gynecologic malignancies harboring PIK3CA mutations. *J Clin Oncol* 2012;**30**:777–82. <https://doi.org/10.1200/JCO.2011.36.1196>.
- Janku F, Kaseb AO, Tsimberidou AM. et al. Identification of novel therapeutic targets in the PI3K/AKT/mTOR pathway in hepatocellular carcinoma using targeted next generation sequencing. *Oncotarget* 2014;**5**:3012–22. <https://doi.org/10.18632/oncotarget.1687>.
- Cáceres A, Jene A, Esko T. et al. Extreme downregulation of chromosome Y and cancer risk in men. *J Natl Cancer Inst* 2020;**112**:913–20. <https://doi.org/10.1093/jnci/djz232>.
- Shi F, Zhang W, Yang Y. et al. Sex disparities of genomic determinants in response to immune checkpoint inhibitors in melanoma. *Front Immunol* 2021;**12**:721409. <https://doi.org/10.3389/fimmu.2021.721409>.
- Ozga M, Nicolet D, Mrózek K. et al. Sex-associated differences in frequencies and prognostic impact of recurrent genetic alterations in adult acute myeloid leukemia (alliance, AMLCG). *Leukemia* 2024;**38**:45–57. <https://doi.org/10.1038/s41375-023-02068-8>.
- Omilian AR, Wei L, Hong C-C. et al. Somatic mutations of triple-negative breast cancer: a comparison between black and white women. *Breast Cancer Res Treat* 2020;**182**:503–9. <https://doi.org/10.1007/s10549-020-05693-4>.
- Yadav S, LaDuca H, Polley EC. et al. Racial and ethnic differences in multigene hereditary cancer panel test results for women with breast cancer. *J Natl Cancer Inst* 2021;**113**:1429–33. <https://doi.org/10.1093/jnci/djaa167>.
- Zhang B, Jia P, Wang J. et al. Integrated analysis of racial disparities in genomic architecture identifies a trans-ancestry prognostic subtype in bladder cancer. *Mol Oncol* 2023;**17**:564–81. <https://doi.org/10.1002/1878-0261.13360>.
- Mahal BA, Alshalalfa M, Kensler KH. et al. Racial differences in genomic profiling of prostate cancer. *N Engl J Med* 2020;**383**:1083–5. <https://doi.org/10.1056/NEJMc2000069>.
- Althubiti MA. Mutation frequencies in endometrial cancer patients of different ethnicities and tumor grades: an analytical study. *Saudi J Med Med Sci* 2019;**7**:16–21. https://doi.org/10.4103/sjmm.sjmm_154_18.
- Jiang Y, Sun A, Zhao Y. et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* 2019;**567**:257–61. <https://doi.org/10.1038/s41586-019-0987-8>.
- Martínez-Ruiz C, Black JRM, Puttick C. et al. Genomic-transcriptomic evolution in lung cancer and metastasis. *Nature* 2023;**616**:543–52. <https://doi.org/10.1038/s41586-023-05706-4>.
- Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol* 1996;**58**:267–88. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005;**67**:301–20. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodol* 2006;**68**:49–67. <https://doi.org/10.1111/j.1467-9868.2005.00532.x>.
- Xindong W, Vipin Kumar J, Quinlan R. et al. Top 10 algorithms in data mining. *Knowl Inf Syst* 2008;**14**:1–37. <https://doi.org/10.1007/s10115-007-0114-2>.

35. Soria D, Garibaldi JM, Ambrogi F. et al. A non-parametric version of the naive Bayes classifier. *Knowl-Based Syst* 2011;**24**:775–84. <https://doi.org/10.1016/j.knsys.2011.02.014>.
36. Harzevili NS, Alizadeh SH. Mixture of latent multinomial naive Bayes classifier. *Appl Soft Comput* 2018;**69**:516–27. <https://doi.org/10.1016/j.asoc.2018.04.020>.
37. Massey Jr FJ. The Kolmogorov–Smirnov test for goodness of fit. *J Am Stat Assoc* 1951;**46**:68–78. <https://doi.org/10.1080/01621459.1951.10500769>.
38. Marko NF, Weil RJ. Non-Gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes. *PLoS One* 2012;**7**:e46935. <https://doi.org/10.1371/journal.pone.0046935>.
39. Zhang M, Liu S, Miao Z. et al. IDEAS: individual level differential expression analysis for single-cell RNA-seq data. *Genome Biol* 2022;**23**:33. <https://doi.org/10.1186/s13059-022-02605-1>.
40. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* 2020;**578**:82–93. <https://doi.org/10.1038/s41586-020-1969-6>.
41. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:1–21. <https://doi.org/10.1186/s13059-014-0550-8>.
42. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40. <https://doi.org/10.1093/bioinformatics/btp616>.
43. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* 2020;**26**:903–9. <https://doi.org/10.1261/rna.074922.120>.
44. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: *Proceedings of the 9th Python in Science Conference* 2010;**7**:92–96.
45. Pedregosa F, Varoquaux G, Gramfort A. et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
46. Li Y, Ge X, Peng F. et al. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biol* 2022;**23**:79. <https://doi.org/10.1186/s13059-022-02648-4>.
47. Duda RO, Hart PE. et al. *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
48. Szklarczyk D, Gable AL, Lyon D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**:D607–13. <https://doi.org/10.1093/nar/gky1131>.
49. Karl Pearson X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London Edinburgh Philos Mag & J Sci* 1900;**50**:157–75. <https://doi.org/10.1080/14786440009463897>.
50. Wilcoxon F. Probability tables for individual comparisons by ranking methods. *Biometrics* 1947;**3**:119–22. <https://doi.org/10.2307/3001946>.
51. Tianzhi W, Erqiang H, Shuangbin X. et al. ClusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* 2021;**2**:100141. <https://doi.org/10.1016/j.xinn.2021.100141>.
52. Martínez-Jiménez F, Muiños F, Sentís I. et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer* 2020;**20**:555–72. <https://doi.org/10.1038/s41568-020-0290-x>.
53. Li J, Lan Z, Liao W. et al. Histone demethylase KDM5D upregulation drives sex differences in colon cancer. *Nature* 2023;**619**:632–9. <https://doi.org/10.1038/s41586-023-06254-7>.
54. Samuels Y, Wang Z, Bardelli A. et al. High frequency of mutations of the PIK3CA gene in human cancers. *Science* 2004;**304**:554–4. <https://doi.org/10.1126/science.1096502>.
55. Thorpe LM, Yuzugullu H, Zhao JJ. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. *Nat Rev Cancer* 2015;**15**:7–24. <https://doi.org/10.1038/nrc3860>.
56. Chen L, Yang L, Yao L. et al. Characterization of PIK3CA and PIK3R1 somatic mutations in Chinese breast cancer patients. *Nat Commun* 2018;**9**:1357. <https://doi.org/10.1038/s41467-018-03867-9>.
57. Maertens O, Johnson B, Hollstein P. et al. Elucidating distinct roles for NF1 in melanomagenesis. *Cancer Discov* 2013;**3**:338–49. <https://doi.org/10.1158/2159-8290.CD-12-0313>.
58. Delon C, Brown KF, Payne NWS. et al. Differences in cancer incidence by broad ethnic group in England, 2013–2017. *Br J Cancer* 2022;**126**:1765–73. <https://doi.org/10.1038/s41416-022-01718-5>.
59. Dressler L, Bortolomeazzi M, Keddar MR. et al. Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the network of cancer genes (NCG) resource. *Genome Biol* 2022;**23**:35. <https://doi.org/10.1186/s13059-022-02607-z>.
60. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;**474**:609–15. <https://doi.org/10.1038/nature10166>.
61. Zhang H, Tomblin G, Weber BL. BRCA1, BRCA2, and DNA damage response: collision or collusion? *Cell* 1998;**92**:433–6. [https://doi.org/10.1016/S0092-8674\(00\)80936-8](https://doi.org/10.1016/S0092-8674(00)80936-8).
62. Dalal H, Dahlgren M, Gladchuk S. et al. Clinical associations of ESR2 (estrogen receptor beta) expression across thousands of primary breast tumors. *Sci Rep* 2022;**12**:4696. <https://doi.org/10.1038/s41598-022-08210-3>.
63. Gulati GS, Sikandar SS, Wesche DJ. et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* 2020;**367**:405–11. <https://doi.org/10.1126/science.aax0249>.
64. Hamelin B, Obradovic MMS, Sethi A. et al. Single-cell analysis reveals inter-and intratumour heterogeneity in metastatic breast cancer. *J Mammary Gland Biol Neoplasia* 2023;**28**:26. <https://doi.org/10.1007/s10911-023-09551-z>.
65. Cheng K, Nair NU, Lee JS. et al. Synthetic lethality across normal tissues is strongly associated with cancer risk, onset, and tumor suppressor specificity. *Sci Adv* 2021;**7**:eabc2100. <https://doi.org/10.1126/sciadv.abc2100>.
66. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 2017;**168**:613–28. <https://doi.org/10.1016/j.cell.2017.01.018>.
67. Dentre SC, Leshchiner I, Haase K. et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* 2021;**184**:2239–2254.e39. <https://doi.org/10.1016/j.cell.2021.03.009>.
68. Church BV, Williams HT, Mar JC. Investigating skewness to understand gene expression heterogeneity in large patient cohorts. *BMC Bioinform* 2019;**20**:1–14. <https://doi.org/10.1186/s12859-019-3252-0>.
69. Kim S. Ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Commun Stat Appl Methods* 2015;**22**:665–74. <https://doi.org/10.5351/CSAM.2015.22.6.665>.
70. Woldrich JM, Mallin K, Ritchey J. et al. Sex differences in renal cell cancer presentation and survival: an analysis of the national cancer database, 1993–2004. *J Urol* 2008;**179**:1709–13. <https://doi.org/10.1016/j.juro.2008.01.024>.
71. Hsu L-H, Chu N-M, Liu C-C. et al. Sex-associated differences in non-small cell lung cancer in the new era: is gender an

- independent prognostic factor? *Lung Cancer* 2009;**66**:262–7. <https://doi.org/10.1016/j.lungcan.2009.01.020>.
72. Webb GI, Boughton JR, Wang Z. Not so naive Bayes: aggregating one-dependence estimators. *Mach Learn* 2005;**58**:5–24. <https://doi.org/10.1007/s10994-005-4258-6>.
 73. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Class* 1999;**10**:61–74. <https://doi.org/10.7551/mitpress/1113.003.0008>.
 74. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–56. Sardinia, Italy: JMLR Workshop and Conference Proceedings, 2010.
 75. García VJ, Gómez-Déniz E, Vázquez-Polo FJ. A new skew generalization of the normal distribution: properties and applications. *Comput Stat Data Anal* 2010;**54**:2021–34. <https://doi.org/10.1016/j.csda.2010.03.003>.
 76. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 2005;**27**:1226–38. <https://doi.org/10.1109/TPAMI.2005.159>.
 77. Jiang L, Cai Z, Wang D. *et al.* Improving tree augmented naive Bayes for class probability estimation. *Knowl-Based Syst* 2012;**26**: 239–45. <https://doi.org/10.1016/j.knosys.2011.08.010>.
 78. Lü J, Wang P. *Modeling and Analysis of Bio-Molecular Networks*. Singapore: Springer, 2020. <https://doi.org/10.1007/978-981-15-9144-0>.