1730–1747 *Nucleic Acids Research, 2020, Vol. 48, No. 4* *Published online 31 December 2019*
doi: 10.1093/nar/gkz1208

Identification of altered biological processes in heterogeneous RNA-sequencing data by discretization of expression profiles

Andrea Lauria[1,2,†], Serena Peirone[2,3,†], Marco Del Giudice[2,4,†], Francesca Priante[2,4], Prabhakar Rajan[5,6], Michele Caselle[3], Salvatore Oliviero[1,2,*] and Matteo Cereda[2,4,*]

[1]Department of Life Science and System Biology, Università degli Studi di Torino, via Accademia Albertina 13, 10123 Turin, Italy, [2]IIGM - Italian Institute for Genomic Medicine, c/o IRCCS, Str. Prov.le 142, km 3.95, Candiolo (TO) 10060, Italy, [3]Department of Physics and INFN, Università degli Studi di Torino, via P.Giuria 1, 10125 Turin, Italy, [4]Candiolo Cancer Institute, FPO - IRCCS, Str. Prov.le 142, km 3.95, Candiolo (TO) 10060, Italy, [5]Centre for Cell and Molecular Biology, Barts Cancer Institute, Cancer Research UK Barts Centre, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK and [6]The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, UK

Received October 16, 2019; Revised December 05, 2019; Editorial Decision December 15, 2019; Accepted December 17, 2019

ABSTRACT

Heterogeneity is a fundamental feature of complex phenotypes. So far, genomic screenings have profiled thousands of samples providing insights into the transcriptome of the cell. However, disentangling the heterogeneity of these transcriptomic Big Data to identify defective biological processes remains challenging. Here we present GSECA, a method exploiting the bimodal behavior of RNA-sequencing gene expression profiles to identify altered gene sets in heterogeneous patient cohorts. Using simulated and experimental RNA-sequencing data sets, we show that GSECA provides higher performances than other available algorithms in detecting truly altered biological processes in large cohorts. Applied to 5941 samples from 14 different cancer types, GSECA correctly identified the alteration of the PI3K/AKT signaling pathway driven by the somatic loss of PTEN and verified the emerging role of PTEN in modulating immune-related processes. In particular, we showed that, in prostate cancer, PTEN loss appears to establish an immunosuppressive tumor microenvironment through the activation of STAT3, and low PTEN expression levels have a detrimental impact on patient disease-free survival. GSECA is available at https://github.com/matteocereda/GSECA.

INTRODUCTION

In recent years, genomic screenings have studied RNA-sequencing (RNA-seq) expression profiles of large cohorts to gain insights into complex phenotypes, including cancer. Despite the abundance of expression data, it remains challenging to identify the biological processes that control disease progression. A major hurdle is the presence of inter-sample heterogeneity (IH), or the variable expression of genes across samples due to genetic, environmental, demographic, and technical factors (1). Furthermore, the admixture of different cell types in the sequenced sample is a well-known source of heterogeneity (2). As the number of samples or the complexity of the phenotype grows, the confounding role of IH in detecting relevant biological information increases (1,3). As a consequence of IH, genes can be expressed at different levels in distinct samples. Specific genes can be activated and repressed in different subpopulations rather than being concordantly expressed in the whole population. Overall, these coordinated heterogeneous changes can result in small expression differences in the whole population that are difficult to detect, especially in large cohorts (4). Moreover, it is well-known that complex phenotypes arise from subtle alterations of distinct genes sharing common functions or involved in the same biological process (i.e. gene sets) in different patients affected by the same condition (5).

In diseases such cancer, heterogeneity strongly impacts on disease progression and drug response (6). Therefore, dissecting the contribution of IH on gene expression becomes crucial to detect defective biological processes and to the therapy management of patients (7). This issue has

*To whom correspondence should be addressed. Tel: +39 0119933969; Email: matteo.cereda@iigm.it
Correspondence my also be addressed to Salvatore Oliviero. Email: salvatore.oliviero@unito.it
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

recently begun to be exploited with single cell analysis (8). Nevertheless, 'bulk' methods of RNA-seq remain the conventional approach to measure gene expression for the advantages of time, cost, and standardized data processing (9). Currently, novel insights on complex phenotypes can be obtained from analyses of the growing public repository of genomic Big Data (10). In this view, the concept of 'pathway' rather than 'single gene' alteration has become widely employed (11). Gene set analysis (GSA) aims at identifying gene sets whose cumulative expression is altered in the phenotype of interest. During the last years, several GSA methods using different statistical tests and null hypothesis formulation have been proposed (11–15). In particular, GSA algorithms can be divided into 'self-contained' and 'competitive' algorithms depending on whether they identify altered gene sets (AGSs) while ignoring or not genes that are outside the gene set of interest, with the former being more powerful than the latter (16).

Most existing GSA methods suffer a few marked limitations (13,16,17). Firstly, GSA algorithms have been designed to handle microarray expression data and subsequently adopted to handle RNA-seq data (11,13). RNA-seq gene expression profiles are characterized by a bimodal behavior reflecting the presence of two major subpopulations of genes in cells (i.e. lowly and highly expressed genes) (18). This behavior is not observable using low-sensitive microarray experiments (19), and to date it has not been taken into account by existing GSA methods. Thus, their application to RNA-seq expression profiles may not be efficient (13). Secondly, GSA methods have been developed to handle experimental conditions in the absence of IH (i.e. altered genes are concordantly activated or repressed in the cohort of interest) (17). As a consequence, biological processes composed of genes that exhibit a significant excess of coordinated variability (i.e. activated or repressed in different subpopulations) cannot be detected by conventional GSA methods (17). Finally, most existing GSA methods have been designed to assess gene expression of case–control studies with limited sample size and, thus, characterized by a negligible IH and high signal-to-noise ratio (13). These limitations become crucial in the analysis of RNA-seq datasets of large-scale screening projects that are characterized by a high IH. Therefore, GSA algorithms that are able to handle IH are needed. To date, few studies carried out a comprehensive analysis of the performance of GSA methods on high-volume RNA-seq datasets (20).

Nowadays, Big Data analysis employs machine learning approaches to assess the contribution of data heterogeneity. It has been recently shown that the division of numerical features into a limited number of non-overlapped intervals (i.e. data discretization, DD) improves the accuracy of such algorithms (21–23). In computational biology, the DD approach has been used to explore gene regulatory networks (24) and, as a pre-processing step, to improve classification accuracy using microarray data (25). Here, we present a Gene Set Enrichment Class Analysis (GSECA) algorithm to identify AGSs in heterogeneous RNA-seq datasets. GSECA implements a sample-specific finite mixture modeling (FMM) approach to assess the bimodal distribution of each RNA-seq profile followed by a model-based DD process to increase the signal-to-noise ratio. Discretized data are then evaluated in a statistical framework to detect AGSs between two groups of samples. We showed that GSECA has the highest sensitivity and specificity in detecting AGSs as compared to other 'state-of-the-art' GSA algorithms in the presence of IH on both simulated and real RNA-seq data. We developed a GSECA as a user-friendly R/Shiny application freely available from GitHub (https://github.com/matteocereda/GSECA).

MATERIALS AND METHODS

Finite mixture modeling of gene expression distributions

To identify the two subpopulations of lowly and highly expressed genes, GSECA models the bimodal distribution of RNA-seq expression profile x of all protein-coding genes of a given sample i as a mixture of two Gaussian probability densities Φ, as previously proposed (18):

$$f(x_i) = \lambda_1\phi(x_1; \mu_1, \sigma_1) + \lambda_2\phi(x_1; \mu_2, \sigma_2) \quad (1)$$

where λ is the mixing proportion, μ and σ are the mean and the standard deviation, respectively (26). To estimate the parameters μ and σ of the two components the method applies the Expectation–Maximization (EM) algorithm (26,27). The algorithm runs iteratively until the maximum likelihood of the parameters of the two components is reached. To ensure a consistent subdivision with the overall expression profile an additional heuristic step was implemented. In particular, GSECA requires the mean of the first component (i.e. highly expressed genes) to be greater than the mean of the second one (i.e. lowly expressed genes). The EM step is repeated until the condition is satisfied. Besides providing an estimate of the Gaussian components, the mixture model calculates the posterior probabilities τ of the component membership of the mixture (26). Thus, GSECA measures the probabilities τ_1 and τ_2 of each gene to belong to the two distributions defined by the two components.

Definition of expression classes

For each sample, genes are considered as (i) not expressed (NE), or not detected, if their expression level (i.e. FPKM) is smaller than 0.01; (ii) lowly expressed (LE) if the probability τ_2 of belonging to the second component of the mixture is greater than 0.9; (iii) highly expressed (HE) if the probability τ_1 of belonging to the first component is greater than 0.9; or (iv) medium expressed (ME) if both the probabilities τ_1 and τ_2 are <0.9. To ensure an adequate distribution of genes among expression classes (ECs), thus a similar degree of statistical power for the subsequent tests performed for all classes, and retain as much information from the original continuous attribute as possible, HE genes are further divided accordingly to the percentiles of the expression level distribution defined by the first Gaussian component (see Supplementary Notes). In particular, for each sample, HE genes were assigned to (i) the first class of high expression (HE1) if their expression level is less than or equal to the 25^{th} percentile of the distribution of HE genes; (ii) the second class of high expression (HE2) if their expression level ranges between the 25^{th} and the 50^{th} percentile; (iii) the third class of high expression (HE3) if their expression level falls

between the 50th and the 75th percentile; or (iv) the fourth class of high expression (HE4) if their expression level is greater than or equal to the 75th percentile.

Statistical analysis of expression classes

After discretizing the gene expression levels into seven expression classes, GSECA implements a statistical framework to detect altered gene sets between the two groups of samples *A* and *B*. First, for each gene *g* and each EC *c*, the number of samples in which *g* is and is not assigned to the class *c, n* and *r*, respectively, are calculated for the two cohorts as follows:

$$\forall g \text{ and } c,\ \ n_{g,c} = \sum_i\ g \in c; \quad r_{g,c} = \sum_i\ g \notin c; \qquad (2)$$

where *i* are the samples in cohorts *A* and *B* (Figure 1).

For each gene set $G = \{g_1,\ldots,g_m\}$, the cumulative number of samples with genes of *G* that are and are not in each expression class across samples of *A* and B, *N* and *R*, respectively, are computed as follows:

$$\forall\ G \text{ and } c,\ \ N_{G,c} = \sum_{g\in G} n_{g,c}; \quad R_{G,c} = \sum_{g\in G} r_{g,c}; \qquad (3)$$

To determine whether cohort *A* is enriched or depleted of genes of a gene set *G* in an EC *c* as compared to cohort *B*, GSECA implements a two-tailed Fisher's Exact test. In particular, GSECA tests the null hypothesis that the cumulative proportions of genes of a gene set in each EC across samples are not different between *A* and *B*:

$$\forall\ G \text{ and } c,\ \ H_o :\ (N_{G,c}; R_{G,c})_A = (N_{G,c}; R_{G,c})_B \qquad (4)$$

As a result, all seven ECs are characterized by a *P*-value representing the alterations (*i.e.* enrichment or depletion) of expression in the gene set. Given the contingency table defined by *N* and *R* for the two cohorts, the algorithm simulates the table under two independent binomial distributions and performs a two-tailed Fisher's Exact test. $R_{G,c}$ is evaluated considering all genes in the gene set that are not in the EC, regardless their class membership. Therefore, all statistical tests perform independent evaluations of the null hypothesis (see Supplementary Notes). In the case of multiple gene sets, the *P*-value of each comparison is corrected for false discoveries using either the Bonferroni or the Benjamini & Hochberg method, respectively, as defined by the user.

Since GSECA tests the overrepresentation of genes in each EC independently from the other ECs (see Supplementary Notes), to quantify the degree of expression perturbation in each gene set *G* between the two cohorts *A* and *B*, the *P*-values of the seven expression classes are combined using the Fisher's method into one goodness-of-fit (X^2) statistic, to obtain the Association Score (AS)(28):

$$\Psi = -2\sum_c \log\left(p\left(c\right)\right) \qquad (5)$$

$$AS\left(G\right) = P_{\text{comb}}\ =\ 1.0 - P\chi^2_{2k}\ \left(\Psi\right) \qquad (6)$$

where Ψ is the combined test statistic and χ^2_{2k} is a Chi-squared distribution with $2k$ degrees of freedom (k = number of ECs), *p* is the *P*-value and *c* is the expression class.

To calculate the significance level of the AS a bootstrapping procedure (random sampling with replacement) is implemented as previously described (29). For 1,000 times, sample labels are shuffled and the AS is calculated for all gene sets. At the end of all iterations, for each gene set, the empirical *P*-value (p_{emp}) is measured as the number of times the AS is smaller than the observed one:

$$p_{\text{emp}}\left(AS_G\right) = \frac{1 + (\sum_i\ AS_{G,i} < AS_G)}{1 + \#\text{iteration}} \qquad (7)$$

Finally, in case the sample sizes differ substantially between cohort *A* and *B*, a bootstrapping procedure (random sampling with replacement) is implemented to measure the success rate (SR). The algorithm down-samples the larger cohort to reach the sample size of the smaller cohort randomly 1,000 times and repeats the analysis at each iteration. At the end of all iterations, for each AS, the SR, or the proportion of significant enrichments (*P*-value < 0.01, two-tailed Fisher's Exact Test) over the total number of comparisons is calculated as previously described (30).

Prostate adenocarcinoma data

Somatic mutations (i.e. single nucleotide variants and small insertion/deletions (InDels)) and RNA sequencing, protein expression and phosphorylation data were downloaded from TCGA Data Matrix portal (Level 3, https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) for 498 prostate adenocarcinoma (PRAD) samples and processed as previously described (6). Briefly, *PTEN* was considered as somatically lost if undergoing homozygous/heterozygous gene deletions, truncating mutations (i.e. stopgain, stoploss, frameshift indels) and damaging mutations. Damaging alterations were defined as missense and splicing (i.e. up to two nucleotides surrounding the splice sites) mutations with predicted damaging effects on the encoded protein. Missense mutations were considered damaging if supported by at least five out of eight function-based scores (SIFT (31), PolyPhen-2 HDIV and HVAR (32), MutationTaster (33), MutationAssessor (34), LTR (35) and FATHMM (36)) or two out of three conservation-based scores (PhyloP (37), GERP++ RS (38), SiPhy (39)). Splicing mutations were predicted as damaging if supported by at least one ensemble score of dbscSNV (40). The copy number status of *PTEN* was assigned as previously reported (41) (see Supplementary Notes). Sample processing was performed using the GeCo++ library (42). The gene set list was composed of 158 manually curated gene sets comprising 4866 human genes from the Kyoto Encyclopedia of Genes and Genomes available from MSigDb35 (version 5, https://software.broadinstitute.org/gsea/msigdb/).

Differential gene expression and gene ontology analysis

Differentially expressed (DE) genes between 75 PTEN-loss and 423 PTEN-wt samples were detected using the R package DESeq2 (43). Briefly, read counts of 19946 genes of each sample were used as input for DESeq2. Genes with read count equal to zero across all samples were removed. An absolute $\log_2$FoldChange ≥ 1 and a false discovery rate ≤ 0.1

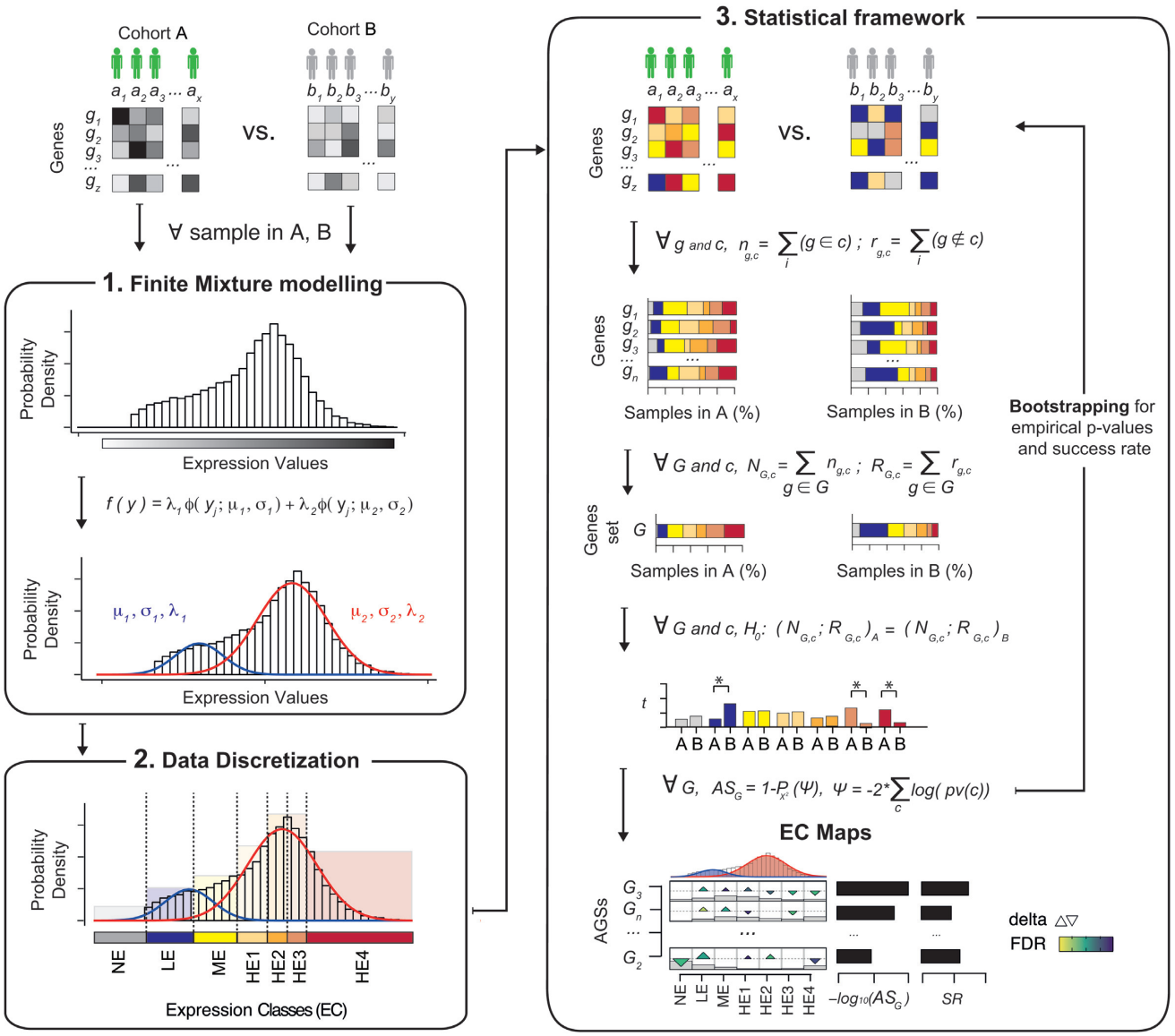

Figure 1. Schematic representation of GSECA algorithm. GSECA requires as input normalized gene expression data of two groups of samples $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_n\}$, and a list of gene sets $G = \{g_1, \ldots, g_n\}$. The algorithm proceeds through three sequential steps: (i) the sample-specific finite mixture modeling of gene expression distribution; (ii) the sample-specific discretization of expression values into seven categorical expression classes and (iii) the statistical identification of altered gene sets (AGSs) obtained by comparing the cumulative proportion of genes of a gene set in each EC between the two cohorts using a Fisher's exact test. The expression perturbation is summarized into an association score (AS), corrected with two bootstrapping procedures for false discoveries (empirical *P*-value) and different sample sizes of the cohorts (success rate, SR). The AGSs are visualized as EC maps. The EC maps display the difference of the cumulative proportion of the genes of a gene set in the seven ECs between the two cohorts as triangles, whose sizes are proportional to such difference. The upper and the lower vertex of the triangles represent enrichment and depletion in cohort A as compared to B, respectively. EC maps depict the proportion *N* of genes in the gene set in each EC as grey bars. GSECA orders AGSs accordingly to their AS, thus obtaining the list of the most altered processes associated with the phenotype of interest.

were used to detect DE genes in PTEN-loss versus PTEN-wt samples. Gene ontology (GO) analysis of DE genes was performed using g:Profiler (http://biit.cs.ut.ee/gprofiler/index.cgi) considering KEGG pathways as gene sets. Statistical results were corrected for multiple comparisons using the native g:SCS method and only gene sets with a corrected *P*-value smaller than 0.05 were retained (44).

PTEN protein–protein interaction network

Proteins interacting with PTEN were retrieved from STRING database (http://string-db.org/). Sources as 'textmining', 'experiment' and 'databases' were used as a type of evidence to measure the interactions between PTEN and other proteins. A minimum interaction score

of 0.9 (i.e. 'highest' confidence) was applied to retrieve the ten best-scoring hits. STRING GO analysis using KEGG pathways was performed with default parameters and gene sets with false discovery rate ≤0.01 were considered as significantly enriched.

Literature-based text mining

Text mining analysis on published journal articles available at the National Center for Biotechnology Information (NCBI) PubMed database was performed using the R package RISMed (https://cran.r-project.org/web/packages/RISmed). For each KEGG gene set, abstracts of published articles were inspected for the co-occurrence of keywords such as 'PTEN' and the gene set nomenclature. For gene sets displaying <50 articles, manual curation of results was performed.

Gene set analysis algorithms

Seven GSA algorithms (i.e. GSVA (13), *Z*-Score (14), PLAGE (15), ssGSEA (12), Globaltest (45), ROAST (46) and GSEA (11)) were used to assess GSECA performances. Implementations of GSVA, *Z*-Score, PLAGE, and ssGSEA methods were available in the R package GSVA. All four methods were run as previously described (13) considering a Poisson kernel to fit RNA-seq expression data. Implementations of Globaltest and ROAST were available from the R/Bioconductor package 'EnrichmentBrowser' (47) and they were run using the *sbea* function with default parameters. The GSEA algorithm was run using GSEA.1.0.R function available from the Broad Institute website (http://www.broadinstitute.org/gsea). Gene sets with corrected *P*-value ≤0.1 were considered significantly associated with the phenotype of interest. To correct for false discoveries due to an unbalanced sample size of the cohorts, the SR of each comparison was measured using bootstrap simulations as previously described (30). For each method, the larger cohort was down-sampled to reach the sample size of the smaller cohort randomly 1,000 times and the analysis was repeated at each iteration. At the end of all iterations, for each comparison, the proportion of significant enrichments (*P*-value < 0.05) over the total comparisons was calculated.

Pancancer dataset

Somatic alterations and transcriptome profiling data were downloaded from the TCGA Data Matrix portal (Level 3, https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) for 9944 and 31 cancer types. For each cancer dataset, samples with somatic loss of PTEN were identified as described for the PRAD cohort. Cancer types with at least 30 samples harbouring the somatic loss of PTEN were retained for further analyses.

Linear regression modeling of AS distributions across cancer types

The distribution of the ASs (i.e. median and inter-quartile range) of 158 KEGG gene sets across cancer types was modeled as a linear regression of six predictors: the number of PTEN-loss and PTEN-wt samples, the median Pearson's correlation coefficient of pairwise comparison of expression profiles in the two cohorts, the statistically significant difference of PTEN expression and of PI3K/AKT signaling pathway cumulative expression (i.e. one- and two-tailed Wilcoxon rank sum *P*-value, respectively). The search for the best subsets of regressors was performed using a branch-and-bound algorithm (48) implemented in the *regsubsets* function in the R 'leaps' package. For models using a different number of variables (i.e. from 1 to 6) the best model in terms of correlation coefficient R^2 was reported. The relative importance of regressors in the linear regression model of six variables was calculated using the function *calc.relimp* in the R 'relaimpo' package (49). This function divides the correlation coefficient R^2 into the contribution of each regressor using the averaging over ordering method (50).

Survival analysis

Clinical data were downloaded from the GDC data portal (https://gdc.cancer.gov/). Disease-free survival time was defined as the interval between the date of treatment and disease progression, as defined by biochemical or clinical recurrence, or until the end of follow-up (51). Disease-free survival analysis was carried out on the PRAD dataset comparing the survival probabilities of PTEN-loss and PTEN-wt samples. The analyses were performed using the *ggsurvplot* function from the R package *survminer*. Kaplan–Meier estimation of the survival probabilities for the two groups of samples (i.e. PTEN-loss and PTEN-wt) was measured and the resulting survival curves were compared using the implemented log-rank test. PTEN TPM optimal cutpoint to separate continuous variables was identified using the *surv_cutpoint* function from the R package *survminer*.

Immune cell composition

Cellular composition of the immune infiltrates for TCGA tumors of 14 cancer types were collected (52). The 'relative number' of immune cells (53) was used as a measure of immune composition. To determine whether the composition of each immune infiltrate was different in PTEN-loss samples as compared to wild-type samples, a Student's *t* test was employed. *P*-values of each comparison are corrected for false discoveries using the Benjamini & Hochberg method. To quantify the degree of perturbation of immune cells, for each cancer type, *P*-values were combined using the Fisher's method into one goodness-of-fit (X^2) statistic (28), referred to as immuno score (IS).

RESULTS

Method overview

We designed GSECA as 'model-based' data discretization (MDD) approach fulfilling both a biological and a 'statistical' requirement. First, we required that the division of expression values into expression classes (ECs) must resemble the presence of two major subpopulations of lowly and highly expressed genes in the cells (18) (i.e. biological requirement). Second, we considered that the discretization process must provide an adequate distribution of genes

among classes and, thus, ensure a similar degree of statistical power for the subsequent tests performed for all ECs (i.e. statistical requirement).

To identify AGSs in a list of gene sets $G = \{G_1, \ldots, G_n\}$ between two cohorts $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_n\}$ of heterogeneous RNA-seq expression profiles, GSECA runs through three sequential steps: (i) the sample-specific analysis of gene expression distribution (Figure 1, Step 1), (ii) the discretization of expression values into ECs (Figure 1, Step 2) and (iii) the statistical identification of AGSs (Figure 1, Step 3). The method is based on a null hypothesis of no over-representation of genes in the gene set in any EC between two cohorts (see Materials and Methods). To increase the signal-to-noise ratio GSECA converts the continuous measurements of expression level into discrete values. For each sample of the cohorts, the algorithm identifies the two subpopulations of lowly and highly expressed genes by fitting a two-component FMM on the normalized expression levels of all protein-coding genes as previously proposed (18) (Figure 1, step 1, see Materials and Methods, Supplementary Notes and Supplementary Figure S1). Using the information derived from the FMM, GSECA defines seven categorical ECs and assigns each gene to the corresponding EC (Figure 1, step 2, see Materials and Methods). Seven is the minimum number of classes that (i) ensures the minimal information loss between the discrete and continuous expression profiles and (ii) provides an adequate distribution of genes among classes (see Supplementary Notes and Supplementary Figures S2 and S3). Finally, GSECA implements a statistical framework to measure the perturbation of each EC between the two cohorts (Figure 1, Step 3). The purpose of GSECA is to evaluate whether the expression pattern of the genes in a gene set shows a significant displacement across the ECs in the samples of interest as compared to controls, thus suggesting a causal relationship between the condition and the phenotype. To quantify the extent of expression perturbation across the ECs for all gene sets, the algorithm combines the significance level of each comparison into an association score (AS) using the Fisher's method (28). To reduce false positives discoveries and correct for a different sample size of the cohorts, GSECA implements two bootstrapping procedures measuring the empirical *P*-value (p_{emp}) and the success rate (SR) of each AS.

We designed GSECA to provide the user with a graphical overview of the variation of expression of each gene set across the seven classes between the two cohorts. GSECA visualizes the AGSs as a heatmap, namely expression class maps (i.e. EC maps), depicting the variation of expression across the seven ECs (Figure 1).

Performance evaluation

To evaluate the performance of GSECA in detecting AGSs we employed simulated and real data sets. For each analysis, we compared GSECA results with those of seven different 'state-of-art' methods (i.e. GSEA (11), GSVA (13), ssGSEA (12), *Z*-Score (14), PLAGE (15), ROAST (46) and Globaltest (45)) that, even if designed to treat microarray data, are widely used in the scientific community to analyze RNA-seq data (Supplementary Table S1).

Type I error rate and statistical power evaluation

In real-life systems, genes are differentially expressed with a certain degree of fold change (FC) and dispersion (i.e. a measure of IH) between two groups of samples (43). GSECA has been developed to dissect the contribution of IH, and thus of dispersion, in large cohorts of samples and detect the truly AGSs. To understand how well the algorithm achieves this aim, we first evaluated its type I error rate and statistical power as compared to the other GSA approaches using simulated RNA-seq data under different parameter settings.

To measure the type I error rate, we generated read counts for N samples and 1000 gene sets of equal size P in the condition of no differential expression as previously proposed (4). Then, for each gene set, we tested the null hypothesis of no difference between the two cohorts (see Supplementary Notes). To examine the effects of sample and gene set sizes, we ran the analysis under different parameter settings of N (60, 150, 300, 500) and P (25, 50, 100, 300), repeating the analysis ten times to obtain more stable results. GSECA resulted in being the most conservative approach, with the lowest type I error rate (average median = 0.002) as compared to the other approaches (average median = 0.05, Figure 2A). The conservativeness of GSECA is due to the conservativeness of the Fisher's Exact test (FET) (54) that are combined into the AS (see Supplementary Notes and Supplementary Figure S4). Each FET depends on to the cumulative number of genes in the gene set in the EC across samples of the cohorts (see Materials and Methods). As the sample size grows, the ability of the test to detect a small variation with high specificity and sensitivity increases (Supplementary Figure S4A and B). As a consequence, combining conservative FET *P*-values using a logarithmic scale (i.e. Fisher's Method) results in small ASs (Supplementary Figure S4C). For this reason, GSECA accounts for false positives better than the other GSA algorithms. Furthermore, GSECA specificity was not influenced by the sample and gene set sizes, remaining constant even in case of large cohorts and large gene sets (Figure 2A).

To assess the statistical power of each algorithm (i.e. the likelihood to detect an AGS when the gene set is altered) we implemented two independent simulations, namely 'FC' and 'dispersion' studies, modelling the contribution of fold and dispersion changes, respectively, in gene expression between two cohorts. In doing so, we first defined three parameters: (i) the proportion of gene sets that contains differentially expressed (DE) genes β, (ii) the percentage of DE genes in each gene set γ and (iii) the FC in gene counts between the two cohorts (iv). Then, we introduced a scaling factor D to control the estimated dispersion in gene counts (see Supplementary Notes). For both studies, we modeled eight conditions where, out of 1000 gene sets, the fractions of truly AGSs β (i.e. true positives) was equal to 5% and 25% and the percentage of DE genes in each gene set was set at 25% and 50% for relatively small and large gene set sizes ($P = \{25,100\}$). For the FC study, we selected FC values ranging between 1.5 and 3 without changes in the estimated dispersion ($D = 1$), thus simulating a differential expression driven by homogeneous changes across samples. Conversely, for the dispersion study, we let D vary between 1.5

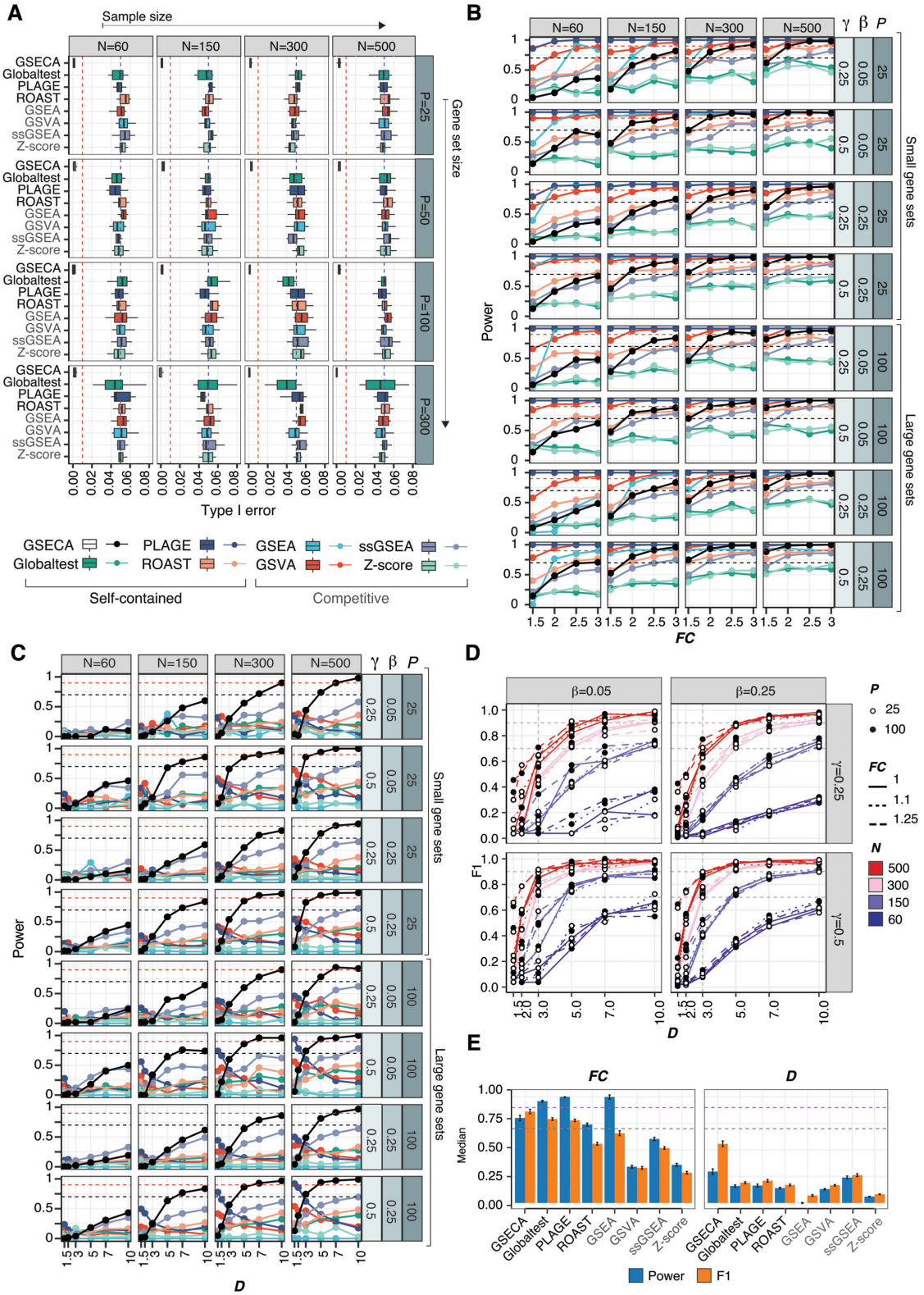

Figure 2. GSA performance evaluation. (**A**) Boxplots depicting the type I error rates for GSA methods evaluated for different settings of sample and gene set sizes on ten replicates. Red and blue dashed lines show the nominal α values of 0.01 and 0.05, respectively. (**B**) Scatter plots depicting the statistical power of each GSA algorithm at the increase of FC between cohorts for different settings of sample size *N*, gene set size *P*, the proportion of gene sets containing differentially expressed genes β, the percentage of DE genes in each gene set γ. (**C**) Scatter plots depicting the statistical power of each GSA algorithm at the increase of dispersion factor *D* at a fixed FC of 1.1 between cohorts for different settings of *N*, *P*, β and γ. (**D**) Scatter plot showing GSECA statistical power at increasing values of dispersion parameter *D* for different settings of *P*, β , γ and FC. (**E**) Bar plots representing the median values of statistical power and *F*1 score measured for all GSA methods in all simulations for the FC and dispersion studies. Gray and purple dashed lines represent values of 0.7 and 0.9, respectively. Black error bars depict standard errors.

and 10 allowing small, or no, FC changes between cohorts (FC = {1,1.1,1.25}), modelling expression changes driven by IH. Under these conditions, we generated read counts for N samples divided into two groups. For each group, we constructed 1000 gene sets composed from P random realizations of negative binomial distribution (4) (see Supplementary Notes). For each gene set, we assessed power by testing the null hypothesis of no differential expression between cohorts for all methods. To account for the different specificity of GSA methods, we measured the F1 score, a performance evaluation metric that provides a harmonic mean of the precision and sensitivity in case of an uneven distribution of true and false positives (i.e. truly AGSs and invariant gene sets, respectively) for all simulations (55).

In presence of FC differences between cohorts (i.e. homogeneous changes across samples), the statistical power of GSECA increased with sample sizes and FC values without being affected by the gene set size and changes in the percentage of DE genes in the gene set (i.e. β and γ, Figure 2B). In particular, GSECA showed a power, or sensitivity, higher than 70% for medium and large sample sizes ($N \geq 150$) under different parameter settings, similar to those of the other self-contained approaches (Figure 2B). Conversely, for small sample size ($N = 60$), other GSA methods show a higher power than GSECA. Overall, we noticed that GSECA predictions showed a better tradeoff between precision and sensitivity ($F1$ score > 0.7) than all other methods even for subtle changes in gene expression for small gene set sizes (i.e. $\beta = 0.05$ and $P = 25$, Supplementary Figure S5A), reflecting the high specificity of GSECA in detecting truly AGSs.

GSECA outperformed other GSA methods in case of IH, and negligible FC, in gene expression between groups (i.e. heterogeneous changes across samples, Figure 2C and Supplementary Figure S5B). In particular, its statistical power grew exponentially with the dispersion parameter (Figure 2C). For small sample sizes (i.e. $N = 60$), none of the algorithms achieved considerable power. GSECA sensitivity increased with sample size whereas the power of all other approaches was almost unaffected with values <70%. Even to a lesser extent, ssGSEA performed similarly to GSECA in handling heterogeneity (Figure 2C). Comparably to GSECA DD approach, ssGSEA brings expression profiles to a common scale collapsing the range of possible gene expression (4). In doing so, ssGSEA reduces the noise of IH (i.e. genes with similar expression levels will have the same rank) increasing its power to detect truly AGSs. GSECA achieved the highest F1 scores, underlining its high sensitivity and specificity in case of heterogeneous gene expression (Supplementary Figure S5B). These results did not considerably change for small variation of FC values or with gene set sizes (Figure 2D and Supplementary Figure S6).

In summary, GSECA has a high sensitivity, proper of self-contained tests (55), of identifying truly AGSs in presence of FC variations between cohorts (Figure 2E, left panel). Most importantly, GSECA is the most powerful GSA approach, among the tested ones, to treat dispersion, and thus IH, in gene expression between phenotypes (Figure 2E, right panel). The results of the simulation studies show that the performances of GSECA are enhanced in case of large cohorts (i.e. $N \geq 150$, Supplementary Figure S7). The statistical power of GSECA increased with sample size as a consequence of the DD, where it is expected that small sample sizes might be not sufficient to estimate the correct distribution of data (56) (see Supplementary Notes and Supplementary Figure S4D). When the IH noise between cohorts is negligible and the cohort size is small, GSECA requires strong FC differences to reach adequate power. GSECA showed the best performance to identify AGSs between heterogeneous cohorts even in presence of overlapping gene sets (Supplementary Notes and Supplementary Figure S8).

Identification of AGSs in PTEN loss prostate adenocarcinomas

To evaluate the performance of GSECA on real data we simulated a condition where known deregulation of a biological process was expected. In particular, a frequently occurring event in prostate cancer is the loss of the tumor suppressor PTEN (57,58) that results in the alteration of the PI3K/AKT signaling pathway (59) and promotes oncogenic programs (60). Among the first ten top-ranked primary interactors of PTEN in the STRING protein-protein interaction (PPI) network (61), nine genes are involved in the PI3K/AKT signaling pathway accordingly to the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Supplementary Figure S9A). Indeed, gene ontology based on the STRING PPI network revealed the enrichment for genes in PI3K/AKT signaling pathway (Supplementary Table S2).

In view of the above, we hypothesized that stratifying human prostate adenocarcinomas (PRADs) accordingly to the somatic loss of PTEN could reveal the altered modulation of the PI3K/AKT signaling pathway. To test this hypothesis, we collected genomic data of 498 PRAD samples available from TCGA and divided them into PTEN loss (PTEN-loss) and wild-type (PTEN-wt) tumors accordingly to somatic mutations and/or copy number alterations in PTEN as previously described (6) (see Supplementary Notes, Supplementary Figure S9B). Using RNA-seq and protein data we measured a significant (i) lower expression of PTEN, (ii) altered modulation of PI3K/AKT genes and (iii) higher phosphorylation level of AKT1 in PTEN-loss tumors compared to PTEN-wt ones (Supplementary Figure S9C-E, P-value < 0.05, two-tailed ranked Wilcoxon test). These results show that the sample stratification led to a dataset characterized by a significantly altered regulation of PI3K/AKT signaling pathway.

We next evaluated the level of IH of the cohorts reflected in the RNA-seq data. Using correlation analyses we found that on average 64% of samples had a low similarity of expression patterns with the others (Pearson's Correlation Coefficient < 0.75, Supplementary Notes, Supplementary Figure S9F and G), confirming the presence of IH in the dataset. To further characterize IH, we measured the FC and dispersion of FPKM values for 19663 protein-coding genes between PTEN-loss and PTEN-wt samples (Figure 3A). We found that 56% of genes showed a reduced expression upon PTEN loss, which was reflected in the higher number of down-regulated genes ($n = 631$) than up-regulated ones ($n = 325$, see Materials and Methods).

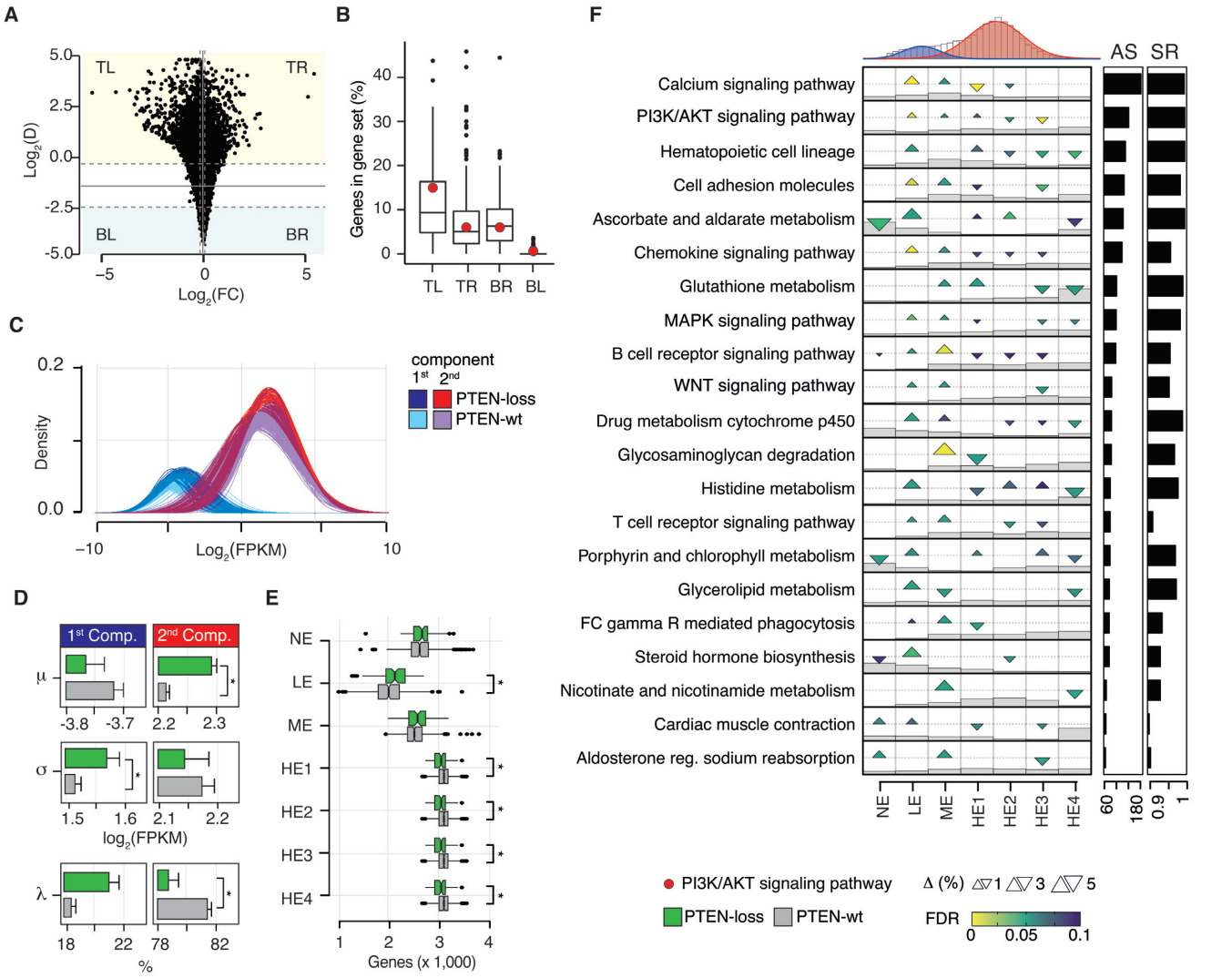

Figure 3. Identification of AGS in PRAD PTEN-loss samples. (**A**) Scatter plot showing the $\log_2$ fold change (FC) and dispersion (*D*) values of all genes between PTEN-loss and PTEN-wt samples. Grey lines represent the median values of FC and *D*. Dashed grey lines show the 25th and 75th percentile of the FC and *D* distributions and define four regions of expression changes (TL = top-left; TR = top-right; BL = bottom-left; BR = bottom-right). (**B**) Boxplots depicting the percentage of genes of each KEGG gene set in the four regions of expression changes. Red dots represent genes in PI3K/AKT signaling pathway. (**C**) Kernel density distributions of FPKM values for PTEN-loss and PTEN-wt samples. (**D**) Comparison of the component parameters (i.e. mean μ, standard deviation σ and mixing proportion λ) defined by the FMM between PTEN-loss and PTEN-wt samples. Statistical tests with a *P*-value <0.05 are considered as significant (*, Student's *t*-test). (**E**) Boxplot distributions of genes in each EC for PTEN-loss and PTEN-wt samples. Statistical tests with a Bonferroni adjusted *P*-value <0.05 are considered as significant (*, two-tailed rank sum Wilcoxon test). (**F**) EC map for the AGSs identified by GSECA in PTEN-loss prostate adenocarcinoma.

Furthermore, out of 4830 genes with a high level of dispersion (i.e. $\geq 75^{th}$ percentile of the dispersion distribution), 29% were activated (i.e. FC $\geq 75^{th}$ percentile of the FC distribution) and 48% repressed (i.e. FC $\leq 25^{th}$ percentile of the FC distribution). These findings highlight a general reduction of expression characterized by IH in the PTEN-loss cohort. This might be a possible consequence of the role of PTEN in regulating basal transcription through histones and chromatin remodeling (62).

To assess the effect of PTEN loss on cellular processes, we applied GSECA on a list of 158 KEGG gene sets (Supplementary Table S3). We found that the fraction of genes in the gene sets reflected the landscape of FC and dispersion values of the cohorts, with the highest proportion of genes being repressed and dispersed (Figure 3B). In particular, 15% of genes in PI3K/AKT signaling pathway were repressed and highly dispersed, suggesting variable repression of PI3K/AKT genes across patients upon the somatic loss of PTEN (Figure 3B).

For each sample, GSECA implemented the FMM approach to model the normal distribution of lowly and highly expressed genes, referred to as 'first' and 'second' components, respectively. As expected, the kernel density distributions (KDDs) of the FPKM values displayed the bimodal profile underlining the presence of the two subpopulations of expressed genes (18) and visually presented different profiles reflecting the presence of outlier samples in the cohorts (Figure 3C).

To assess whether the FMM depicted PRAD IH, we compared the component parameters (i.e. μ, σ, and λ) between the cohorts. Even if showing the same μ, the average σ of the first component (i.e. lowly expressed genes) was significantly higher in PTEN-loss samples as compared to PTEN-wt ones (P-value = 0.031, two-tailed Student's t-test, Figure 3D, left panels), suggesting that PTEN-loss are more heterogeneous at low levels of expression (i.e. distinct lowly expressed genes in different patients) as compared to PTEN-wt samples. Inspecting the second component (i.e. highly expressed genes), we found that, on average, μ was significantly higher in PTEN-loss as compared to PTEN-wt samples (P-value = 0.003, two-tailed Student's t-test) whereas σ did not differ between the cohorts (Figure 3D, right panels). These results confirmed that IH is less pronounced for high expressed genes than lowly expressed ones in PTEN-loss tumors. Finally, comparing the mixing proportions λ we found, on average, a significantly higher number of genes assigned to the first component (i.e. lowly expressed) in PTEN-loss as compared to PTEN-wt samples (P-value = 0.001, two-tailed Student's t-test, Figure 3D, left bottom panel). These results indicate that the FMM fully captures the features of the expression landscape of the cohorts resulting in a higher degree of IH at low expression levels upon PTEN loss.

Next, GSECA applied the DD process and clustered genes into the seven ECs for each sample of the two cohorts. We compared the distribution of the proportion of genes in each EC and found a significant increase and reduction of genes in the LE class (Bonferroni adjusted P-value = 0.016, two-sided Wilcoxon test) and in the four HE classes, respectively, in PTEN-loss samples as compared to wild-type ones (Bonferroni adjusted P-value = 0.021, two-sided Wilcoxon test, Figure 3E), reflecting the differences detected by the FMM. This result confirmed that the DD process preserves the structure of the expression dataset.

Then, GSECA compared the fraction of genes in each of the seven classes between PTEN-loss and PTEN-wt samples in the 158 KEGG gene sets. For each gene set, we determined the enrichment or depletion of samples with genes in each EC and corrected this for multiple tests. Furthermore, for all gene sets GSECA provided the AS, the associated empirical P-value to avoid false discoveries, and the SR controlling for the different sizes of the two cohorts (see Materials and Methods). We found that 21 out of 158 KEGG gene set were significantly altered in PTEN-loss as compared to wild-type samples (AS $\leq$ 0.01, $p_{emp} \leq$ 0.001 and SR $\geq$ 0.9, Figure 3F, Supplementary Table S4). Among these gene sets, GSECA identified the PI3K/AKT signaling pathway as the second top-ranked AGSs, showing a significant increase in the number of samples expressing genes in the LE, ME, and HE1 classes and a significant decrease in the HE2 and HE3 classes (FDR $<$ 0.1, Figure 3F, Supplementary Table S4), supporting the presence of high IH for PIK/AKT genes at low level of expression (Figure 3B). Among the remaining AGSs, GSECA identified five gene sets of signal transduction (i.e. calcium signaling, cytokine-cytokine receptor interaction, cell adhesion molecules (CAMs), MAPK and WNT signaling pathway) that are tightly connected with PI3K/AKT signaling pathway. In particular, PTEN silencing, and the subsequent alteration of PI3K/AKT pathway, impairs calcium signaling (63), alters epithelial CAMs and focal adhesion gene expression in prostate (64), alters MAPK (65) and WNT signaling cascade (66). Furthermore, GSECA detected the alteration of five immune-related processes (i.e. hematopoietic cell lineage, chemokine signaling pathway, B and T cell receptor signaling, FC gamma R mediated phagocytosis), supporting the role of PTEN in regulating the proliferation and differentiation of hematopoietic stem cell (67), controlling signaling and homeostasis in both B and T cells (68,69), and inhibiting FC gamma receptor signaling (70), as well as the role of PI3K/AKT pathway in the regulation of chemokine signaling during prostate tumorigenesis (71). Finally, GSECA highlighted the alteration of nine metabolic pathways (Figure 3F and Supplementary Table S4), underlining the contribution of PTEN in metabolism control (72). GSECA also identified the alteration of cardiac muscle contraction and aldosterone-regulated sodium reabsorption gene sets (Figure 3F and Supplementary Table S4). It is worth noting that the down-regulation of PTEN decreases heart muscle contractility (73) and the activation of PI3K/AKT pathway might be responsible for the alteration of aldosterone-mediated sodium transport in epithelial cells (74).

These results indicate that the FM modelling of RNA-seq expression profiles, the sample-specific DD process and the statistical framework implemented in GSECA can successfully identify known altered biological processes (i.e. PI3K/AKT signaling pathway) in a phenotype of interest (i.e. PTEN somatic loss) considering datasets characterized by IH.

Comparison with available GSA algorithms

We next compared the performances of GSECA in detecting AGSs in PRAD samples upon the somatic loss of PTEN with those of the other GSA methods. Since these methods were not designed to work with an unbalanced sample size between two cohorts (17), we measured the success rate of each comparison using bootstrap simulations as previously described (30) (see Materials and Methods). We found that Z-Score, PLAGE, ssGSEA identified an average of 20 AGSs, comparably to GSECA, whereas GSVA identified 41 AGSs, and ROAST and GSEA detected only one AGS (adjusted P-value $<$ 0.1, SR $>$ 0.9, Figure 4A). To assess the concordance of results among methods, we measured the proportion of shared AGSs over the total number of unique AGSs between any couple of algorithms (i.e. Jaccard Coefficient, JC). Overall, the similarity of results among methods was low (mean JC = 12%). GSECA identified four AGSs that were not detected by any other methods (i.e. T and B cell receptor signaling pathway and steroid hormone biosynthesis) and showed the higher concordance of results with ssGSEA (JC = 19%, Supplementary Figure S10). The weak concordance of results among different GSA algorithms has been previously reported as a consequence of the distinct statistical assumptions of each method (17). Thus, the number and the type of algorithm-specific AGSs might reflect the ability of a method to handle the IH of a complex dataset.

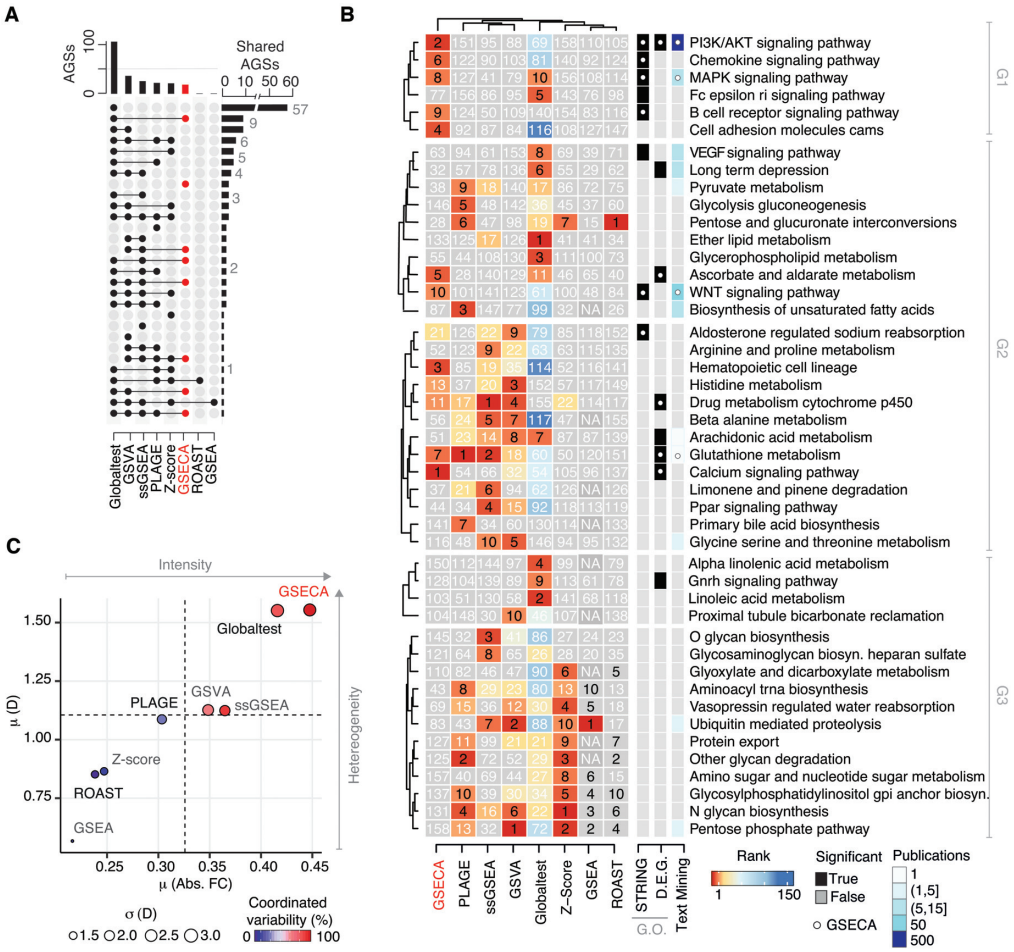

Figure 4. Performance evaluation of GSA algorithms on the PRAD dataset. (**A**) Overlap of AGSs in PRAD PTEN-loss samples identified by the GSA algorithms. GSECA results are reported in red. (**B**) Hierarchical clustering of the first ten top-ranked AGSs in PRAD PTEN-loss detected by GSA algorithms. Each cell reports the rank of the gene set of a specific method. The ranks of the top ten ranked gene sets are reported in black. Annotation heatmap (right) depicts gene sets identified by GO analysis performed considering the STRING PPI network and differentially expressed genes (i.e. DEG) in black and (ii) evidence coming from literature text mining in color key of blues. (**C**) Scatter plot of the mean absolute FC (Abs. FC), and *D* averaged on the 20 top-ranked gene sets detected by each method. Dot size represents the average standard deviation (σ) of *D* for the 20 top-ranked gene sets. Color key depicts the percentage of the 20 top-ranked gene sets that contain both activated and repressed gene sets, namely coordinated variability.

To evaluate the biological reliability of the information provided by each method, we compared GSA results with those of orthogonal approaches. In doing so, we first selected the ten top-ranked AGSs identified by each algorithm and hierarchically clustered their rankings. Next, we performed gene ontology (GO) analyses using the STRING PPI network and the differentially expressed genes in PTEN-loss as compared to PTEN-wt tumors. Finally, we integrated GSA and GO results with a text mining analysis of published articles exploring the connection of PTEN with the selected AGSs (see Materials and Methods). The hierarchical clustering of AGSs highlighted the presence of five groups composed by gene sets that were (i) mainly identified by GSECA and Globaltest (G1); (ii) generally detected by different methods (G2) and (iii) exclusively marked by methods other than GSECA (G3, Figure 4B). Remarkably, GSECA was the only algorithm able to detect, the alteration of PI3K/AKT signaling pathway as a top-ranking result. We found that G1 contains 75% of gene sets enriched from the STRING PPI gene ontology, suggesting the highest ability of GSECA in detecting gene sets that are functionally related to the somatic loss of PTEN as compared to the other methods (Figure 4B). Considering differentially expressed genes, we found that five out of the eight gene sets enriched for significantly up- and down-regulated genes were detected by GSECA (Figure 4B, column 'DEG'). The GO analysis revealed enrichment of PI3K/AKT genes only when we considered significantly activated and repressed genes (Supplementary Table S5). The fact that GSECA was the only algorithm able to identify this gene set as top-ranked AGSs suggests its ability in identifying processes where genes are significantly altered in both directions rather than being either activated or repressed. Finally, four out of the ten top-ranked GSECA AGSs showed published evidence ($\geq$9) for the interaction of PTEN with the gene set, with the highest number of articles (n = 499) supporting PTEN regulation of PI3K/AKT signaling pathway (Figure 4B, right column and Supplementary Table S5). Conversely, we found few published records (mean = 4) supporting the interaction of PTEN

with AGSs that were detected by other methods (Figure 4B and Supplementary Table S5). In particular, AGSs that were exclusively marked by methods other than GSECA did not show any supports from GO and text mining results, with the only exception of the pentose phosphate pathway (Figure 4B). Overall, GSECA was the only algorithm able to identify the altered modulation of PI3K/AKT pathway and of other known and functionally related processes in a real dataset characterized by IH. We further evaluated the degree of these algorithms to handle the heterogeneity of PRAD dataset assessing (i) the minimum number of samples required to detect alterations of the ten top-ranked GSECA results and (ii) how the stratification of samples according to PTEN expression levels affected the results (see Supplementary Notes). Using simulation studies, we found that GSECA achieves the best results in identifying the alteration of PI3K/AKT genes at the increase of IH (i.e. smaller cohort sizes or less stringent PTEN stratification, Supplementary Figures S11 and S12 and Table S6).

To gain further insights into the reasons why GSECA outperformed the other algorithms in detecting the alteration of PI3K/AKT signaling pathway upon PTEN loss, we analyzed the gene expression levels of the 20 top-ranked gene sets of each method independently from their statistical significance. As mentioned above, a consistent fraction of PI3K/AKT genes is down-regulated in a highly variable manner across samples upon PTEN somatic loss (Figure 3B and Supplementary Figure S13A). Hence, to detect the alteration of this gene set, GSA methods must handle 'coordinated' expression changes (i.e. activation or repression) of distinct genes in different samples even if they can also result in small FC differences in the whole population. Therefore, we first inspected the distribution of expression levels of genes in each gene set across PRAD samples (Supplementary Figure S13B). GSECA, ssGSEA and Globaltest detected gene sets predominantly composed by genes expressed at both low and high levels (see Supplementary Notes and Supplementary Figure S13B). Conversely, the other algorithms detected gene sets predominately composed of highly expressed genes, suggesting that they might be not effective in detecting changes of lowly expressed genes. This held particularly true for GSEA whose KDDs showed the strongest shift towards high expression levels (Supplementary Figure S13B).

To corroborate these findings, we assessed four parameters (i.e. FPKM values, FC, absolute FC and dispersion, see Supplementary Notes) measuring the range, direction, intensity, and IH of gene expression captured as altered by each method. In particular, we evaluated the mean μ and standard deviation σ of each parameter across PRAD samples and, then, averaged results across genes (Supplementary Figure S13C). Overall, GSECA showed the best results in handling gene sets characterized by expression changes of groups of genes that are more intensively activated or repressed (i.e. direction and intensity, Supplementary Figure S14) at all levels (i.e. range) in a heterogeneous manner across samples (i.e. IH, Figure 4C). For this reason, GSECA was able to detect the altered modulation of PI3K/AKT signaling pathway that is composed of genes that are expressed at different levels (*i.e.* low and high) and are distinctly activated or repressed in different samples upon PTEN loss (Supplementary Figure S13B and C). It is worth noting that Globaltest performed similarly to GSECA. However, the highest type I error (i.e. 119 AGSs out of 158 gene sets, Figure 4) confers less confidence to the results (i.e. PI3K/AKT signaling pathway rank = 69, Figure 4B). Remarkably, our analyses show that other methods, particularly GSEA, cannot handle datasets characterized by high IH (Figure 4 and Supplementary Figure S13C).

These results show that GSECA can detect functionally relevant altered biological processes under a phenotype of interest when considering more heterogeneous cohorts in contrast to other available methods.

The somatic loss of PTEN impacts on immune-related processes

Somatic inactivation of PTEN occurs in a wide range of human cancers with various effects on each tissue (75). For this reason, we employed GSECA to perform a comprehensive analysis of biological processes that are altered upon the loss of PTEN across cancer types. In particular, we collected genomic data of 9944 samples of 31 cancer type available from TCGA (Supplementary Table S7) and stratified them with respect to somatic alterations in PTEN as described for the PRAD cohort. Together with PRAD, we retained for further analyses 13 cancer types for which we could identify at least 30 samples with somatic alteration of PTEN (Supplementary Table S7). As expected, we measured a significant decrease in PTEN expression levels in PTEN-loss samples as compared to wild-type ones (P-value < 0.05, one-tailed Wilcoxon Rank sum test, Supplementary Figure S15A), leading to a significant alteration of expression pattern of the PI3K/AKT signaling pathway in corpus endometrial carcinoma (UCEC), low grade glioma (LGG), skin cutaneous melanoma (SKCM) and sarcoma (SARC) (P-value < 0.05, two-tailed Wilcoxon Rank sum test, Supplementary Figure S15B). Using GSECA, we discretized the gene expression levels into the ECs and found a significant alteration of the number of genes in the ECs between PTEN-loss and PTEN-wt samples in UCEC, LGG, SKCM, lung squamous carcinoma (LUSC) and kidney chromophobe (KICH, Supplementary Figures S16 and S17), suggesting a possible regulation of transcriptional programs driven by PTEN in these tissues. We next compared the fraction of genes in the ECs between PTEN-loss and PTEN-wt samples in the list of 158 KEGG gene sets looking for significant differences. We found that 10 out of 13 cancer types showed at least one AGS (AS $\leq$ 0.05, $p_{emp} \leq 0.05$ and SR ≥ 0.7, Supplementary Table S8). Importantly, six cancer types (i.e. UCEC, LGG, HNSC, SARC, PRAD and BRCA) showed the significant alteration of the PI3K/AKT signaling pathway expression pattern (Figure 5A). Indeed, GSECA AS showed a significant positive correlation with the extent of PI3K/AKT signaling pathway alteration, which was measured as the statistical difference in the cumulative expression of PI3K/AKT related genes in PTEN-loss tumors as compared to wild-type samples (two-tailed Wilcoxon Rank sum test, Figure 5B). To further assess whether the AS exploits the alteration of PI3K/AKT signaling pathway, we employed a linear regression approach. In particular, we modeled the AS distri-

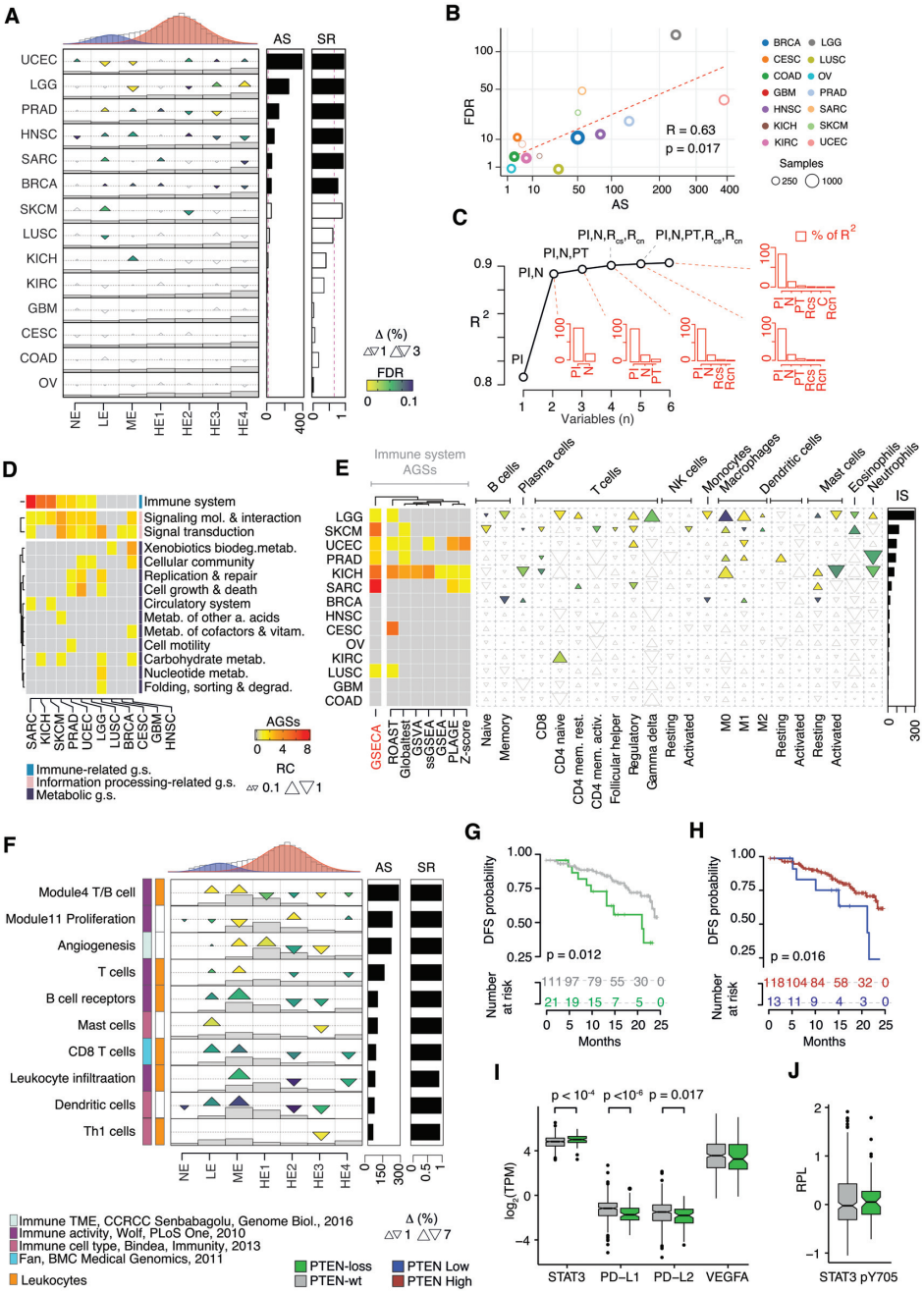

Figure 5. Pan-cancer analysis of PTEN somatic loss. (**A**) GSECA EC map showing the pan-cancer alteration of PI3K/AKT signaling pathway as a consequence of the somatic loss of PTEN. (**B**) Scatter plot showing the correlation of GSECA AS (*i.e.* -10*log10(AS)) and the alteration of PI3K/AKT signaling pathway in PTEN-loss as compared to PTEN-wt tumors measured in terms of the adjusted *P*-value (*i.e.* -10*log10(FDR)) across cancer types. The size of the colored circles shows the number of samples, while the inner white circles the number of PTEN loss samples. (**C**) Coefficients of determination (R^2) of the linear regression model at the increasing of model complexity (i.e. the number of regressors in the model). PI=PI3K/AKT signaling pathway alteration; PT=PTEN downregulation; *N*=number of samples; R_{CS}=correlation of PTEN-loss samples; R_{CN}=correlation of PTEN-wt samples. Bar plots in red show the **r**elative importance of each predictor to the R^2 measured by the linear regression model of all variables. (**D**) Heatmap showing the altered classes of gene sets across cancer types. Classes are defined accordingly to the KEGG category. Each cell reports the number of AGSs. The annotation heatmap indicates the KEGG superclass of biological processes. (**E**) Heatmap on the left panel shows the number of immune-related gene sets that are altered upon the loss of PTEN across cancer types accordingly to GSECA and the other GSA methods. On the right panel, EC map-like heatmap depicts the statistically significant alteration of the immune cell population across cancer types. The size of triangles the relative change of the percentage of tumor immune infiltrates between PTEN-loss and wild-type samples. Upper/lower vertexes of the triangles represent the increase/decrease of immune cells in PTEN-loss samples as compared to PTEN-wt tumors. The bar plot reports the IS for each cancer type. (**F**) GSECA EC map showing the altered immune expression signatures as a consequence of the somatic loss of PTEN in PRAD. (**G**). Disease-free survival (DFS) Kaplan-Mayer curves for PTEN-loss and PTEN-wt patients. (**H**). DSF Kaplan-Mayer curves measured stratifying PRAD patients on the optimal PTEN expression level (i.e. TPM=3.56, maximally selected rank statistics=2.34) within two years from the initial treatment. (**I**) Boxplots showing expression distributions of PTEN normalized expression levels for PTEN-loss and PTEN-wt samples of four immune-response related genes. (**J**) Boxplot distributions of the relative level of STAT3 phosphorylation for PTEN-loss and PTEN-wt PRAD samples.

bution across cancer types as a function of the alteration of both PTEN and PI3K/AKT signaling pathway expression levels, the number of PTEN-loss samples, and the correlation of expression profiles in PTEN-loss and PTEN-wt tumors (see Materials and Methods). We then performed an exhaustive search for the best subsets of variables for predicting the variability of the AS distributions using a branch-and-bound algorithm (48) (see Materials and Methods). We found that the alteration of PI3K/AKT signaling pathway gave the best fitting of the AS distributions in terms of coefficient of determination when using one predictor ($R^2 = 0.81$, Figure 5C). Increasing the model complexity led to a closer fitting between the AS distributions and the predictors (Figure 5C). We next measured the relative importance of each regressor in the linear model using the averaging over ordering method (50) (see Materials and Methods). We found that even considering all variables in the model the alteration of expression of the PI3K/AKT signaling cascade was the most critical regressor accounting for 80% of the explained variance by the model (P-value = 0.009, Figure 5C). These results indicate that GSECA AS recapitulate the extent of PI3K/AKT cascade alteration. Furthermore, GSECA identified PI3K/AKT signaling pathway as altered in two cancer types (i.e. UCEC and LGG) for which the alteration of PI3K/AKT signature is known to impact on patient survival in positive and negative way (76,77). The survival analysis based on PTEN loss stratification in UCEC and LGG confirmed these results, reinforcing the robustness of GSECA prediction (Supplementary Notes and Supplementary Figure S18A).

To gain functional insights on the cancer-specific regulation of PTEN, we next inspected the 10 top-ranked AGSs in each cancer type and hierarchically clustered them at an intermediate KEGG gene set category level (78). We found that (i) metabolic processes were explicitly altered in distinct cancer types, (ii) information-related processes were altered among different tumor types, and, importantly, (iii) immune system gene sets were altered in the majority of tissues (Figure 5D). In particular, SARC, KICH and SKCM showed the highest number of immune-related AGSs, being hematopoietic cell lineage, chemokine and T cell receptor signaling pathways the most altered gene sets across cancer types (Supplementary Table S9). These results highlight the association between the loss of PTEN and the alteration of immune cell infiltrates, which has been recently noted (2). To verify the accuracy of the results of GSECA, we performed two different analyses. First, we compared the results of the other GSA methods on the immune-related gene sets. Second, we evaluated the changes in the tumor immune microenvironment (TIME) upon the loss of PTEN. For the latter analysis, we collected information about the cellular composition of immune infiltrates for TGCA tumors of 14 cancer types (52) and statistically measured the differences in the composition of 22 distinct immune cell types between PTEN-loss and PTEN-wt samples. Finally, to provide the degree of alteration of the immune cell population we combined the significance level of each comparison into an immune score (IS) using the Fisher's Method (see Materials and Methods, Supplementary Figure S18B and Table S10). Compared to the other GSA methods, GSECA detected the highest number of cancer types with a significant alteration of immune cell fractions (Figure 5E, left panel). Moreover, GSECA showed the highest positive correlation between the number of immune-related AGSs and the IS across GSA methods (Pearson's correlation coefficient $R = 0.77$, P-value = 0.003, Figure 5E, right panel). In particular, the AS resulted significantly positively correlated with the IS, highlighting the accuracy of GSECA results (Supplementary Figure S18C). Together these results indicate that GSECA was the most robust approach to highlight the link between PTEN loss and alteration of immune regulation by detecting the highest number of immune-related AGSs in the vast majority of cancer types with statistically significant changes in TIME composition.

Emerging evidence has suggested that PTEN loss is an immunosuppressive event in prostate tumors (79). However, the connection between PTEN and the immune system is complex and involves both pro- and anti-tumorigenic immune responses depending on the cellular phenotype and the TIME (80). To assess the general applicability of GSECA we finally sought to investigate the impact of PTEN loss on TIME of PRAD samples. In doing so, we ran GSECA on a collection of 102 expression signatures representative of different immune cell activities, states, and modes in tumor tissues (52). We found that 15 immune signatures were significantly altered upon PTEN loss (AS $\leq$ 0.05, $p_{emp} \leq 0.01$ and SR ≥ 0.7, Supplementary Table S11). Six of the top ten AGSs characterized the state and activity of T and B cells, showing a general reduction of gene expression in the highly expressed classes, and a reciprocal increase of genes in lowly expressed classed (Figure 5F). These observations are consistent with previous data suggesting that PTEN loss prostate cancers are non-T cell inflamed, or 'cold', tumors (81). In particular, GSECA identified the decreased expression of genes representative of CD8 T cells, which was supported by the results of the TIME analysis (Figure 5D), as previously reported (79).

Interestingly, the two top-ranked AGSs contained markers of lymphocyte activation (i.e. Module4 T/B cells) and cell proliferation (i.e. Module11 Proliferation), respectively (82). The combination of the down-regulation of the T/B cell module and the upregulation of the proliferation module has been strongly associated with decreased disease-free survival (DFS) in breast cancer patients (82). Since GSECA identified this same pattern, showing a reduction of genes in HE classes for the T/B cell module and an increase for the proliferation module (Figure 5F), we assessed the impact of PTEN loss on DFS in prostate cancer patients. Using DFS data for 355 PRAD patients available from TCGA, we observed a statistically significant difference in time to disease progression in the first 24 months from the treatment between patients with PTEN loss and wild-type (Figure 5G and Supplementary Figure S18D). These data confirm the detrimental impact of PTEN loss on prostate cancer disease phenotype. Furthermore, since PTEN status determination impact on therapy management of prostate cancer patients (80), we wondered whether absolute PTEN expression levels could be prognostic of a shorter DFS. Using the maximally selected rank statistics approach (83), we found that patients with PTEN expression levels lower than 3.58 TPM had a statistically significant shorter DFS time in the first two years (Figure 5H), as well as three years (Supplemen-

tary Figure S18E), from the initial treatment. These results highlight that not only the genomic status but also the absolute expression levels of PTEN are associated with poor outcomes in patients with prostate cancer.

It has been shown that in *Pten*-null mice the activation of the Stat3 establishes an immunosuppressive TIME that contributes to tumor growth and chemoresistance (84). Therefore, to finally validate GSECA results, we compared the expression levels of STAT3 in PRAD PTEN loss tumors as compared to controls. We also evaluated the expression of the inhibitory immune checkpoint molecule PD-L1 and PD-L2 and the immune inhibitor VEGFA (52). We found that PRAD PTEN loss tumors significantly expressed STAT3 at higher levels and PD-L1 and PD-L2 at lower levels than PRAD PTEN wild-type samples (Figure 5I). Moreover, the level of phosphorylation of STAT3 was significantly higher in PRAD PTEN loss tumors as compared to controls (Figure 5J). These data support the establishment of an immunosuppressive TIME in human prostate cancers, which could be driven by the activation of STAT3, and validate the statistically significant associations found by GSECA.

Taken together, these results show the general applicability of GSECA in detecting biological processes that are altered in high-volume heterogeneous data sets, including pathological and physiological conditions other than cancer (Supplementary Notes, Supplementary Figure S19 and Table S12). In particular, GSECA has proved highly accurate in associating the loss of PTEN to the alteration of PI3K/AKT signaling pathway and to the different regulation of immune-related processes across cancer types. In prostate cancer, GSECA detected the detrimental impact of PTEN loss on DFS of patients and the establishment of a 'cold' TIME through the down-regulation of lymphocytes signatures. Hence, our results support the emerging role of PTEN in immune system (2,85,86) and therapy resistance (87–89).

DISCUSSION

In this study, we have developed and evaluated GSECA, a tool to identify altered biological processes from the analysis of high-volume and heterogenous RNA-seq experiment data. Heterogeneity is a fundamental characteristic of information associated with complex traits, which can arise from subtle deregulation of distinct genes in different patients rather than of a single gene (6,11,30). IH affects the ability to detect such modifications in large datasets. Here we explored the importance of IH for the correct identification of biological mechanisms that are relevant for the phenotype of interest and confirm its confounding role in signal detection (90,91).

By exploiting the concept of two major subpopulations of genes expressed by the cell (18), GSECA estimates the IH due to biological and technical conditions. The method employs sample-specific estimates of gene expression distribution in a discretization process that reduces the large set of numerical values into a small list of categorical values. Finally, using these discrete units of measure, GSECA identifies the processes that are significantly altered in the phenotype of interest.

We employed simulated RNA-seq data modelling several conditions of differential gene expression between two cohorts to evaluate the performance of GSECA as compared to other seven 'state-of-art' GSA methods. Overall, GSECA showed a deflated type I error rate and a higher power than the other GSA methods when handling heterogeneous RNA-seq datasets. The algorithm also achieved a substantial sensitivity to detect AGSs in absence of IH, even if not designed to treat this scenario. Most importantly, GSECA displayed the highest F1 score among all methods to detect truly AGSs in the presence of IH in gene expression between samples. To summarize, GSECA can identify a smaller number of AGSs as compared to other GSA methods but with a higher sensitivity. The conservativeness of GSECA allows to avoid the 'overproduction' of significant results (92), and to provide the user a narrowed list of truly altered processes to inspect in details in further analyses. The predictions of GSECA were the most accurate ones when treating heterogeneous samples suggesting that its framework enhances the signal-to-noise ratio, and thus data interpretation. Interestingly, ssGSEA showed similar performance to GSECA in handling heterogeneity. Comparably to GSECA, this method treats each sample individually and collapses gene expression levels to a common scale using ranks (12). This finding confirms that the reduction of the large set of expression levels into a smaller range of values increases the power to detect truly AGSs in the presence of IH.

We used our approach to identify the biologically relevant gene sets that are altered upon the somatic loss of PTEN, and the subsequent alteration of the PI3K/AKT signaling cascade in prostate cancer. The EC maps generated by GSECA correctly detected the alteration of the PI3K/AKT signaling pathway and related signal transduction gene sets, such as calcium signaling (63), epithelial CAMs (64), MAPK (65) and WNT signaling pathways (93). Interestingly, the FMM and the DD approaches captured the heterogeneity among the cohorts revealing a general decreased and widespread gene expression in prostate cancer due to the loss of PTEN that might underline the role of PTEN in regulating basal transcription through histones and chromatin remodeling (62).

The comparative performance analysis of GSA methods in detecting the effect of PTEN silencing shows that GSECA was the only algorithm able to tackle the heterogeneity of the prostate cancer dataset and to reveal the altered modulation of PI3K/AKT signaling pathway. Moreover, GSECA detected the altered regulation of processes where genes directly interact with PTEN and, thus, are influenced by the somatic loss of their interactor. This result indicates the ability of the method to spot functionally related AGSs. Importantly, GSECA highlights the alteration of gene sets composed by genes that are coordinately and heterogeneously modulated rather than being uniformly activated or repressed at different levels (i.e. low and high) in distinct samples, whereas other methods might suffer from this limitation. Together, these results indicate that GSECA boosts the signal-to-noise ratio in heterogeneous datasets enabling the identification of the general mechanisms that are more likely to be altered across samples. In particular, compared to the other GSA algorithms, GSECA requires

a smaller number of heterogeneous samples to detect the AGSs, and it can handle different degrees of IH without affecting the final results.

The pancancer analysis of the effect of PTEN somatic loss generated a comprehensive assessment of its regulation across tissues. PTEN critically interconnects the canonical PI3K/AKT and the RAS/MEK/ERK pathway, which are the two dominant tumorigenic gene sets controlling cell survival and proliferation (75). Our data shows that the impact of PTEN silencing on cellular program regulation is proportional to the impaired modulation of the PI3K/AKT signaling cascade, with the stronger effect of gliomas, endometrial, head and neck, breast carcinomas, melanomas, and sarcomas. GSECA revealed a tissue-specific control of PTEN on metabolic processes, whereas information-related processes, such as signal transduction, are more uniformly affected across tissues. Most importantly, GSECA correctly highlighted the role of PTEN in controlling immune-related processes in the majority of cancer types, particularly in those showing a significant alteration of the TIME composition. These data support the importance of PTEN in modulating the immune system (85) and therapy resistance (89). Recently, it has been shown that the loss of PTEN impacts on T cell lineage stability (86), inhibits T cell-mediated tumor killing and trafficking in melanomas (88,94) and promotes resistance to T cell-mediated immunotherapy in uterine cancers (87). Using additional immune expression signatures, GSECA correctly highlights the immunosuppressive TIME of PTEN-loss prostate tumors (80), which could be driven by the significant activation of STAT3. Furthermore, GSECA results were pivotal to show the shorter of disease-free survival of these patients and to underline the biomarker potential of PTEN expression levels. These results validate previous findings in prostate mouse models (84), melanoma (88), breast (82) and provide indications that might be important for the clinical management of prostate cancer patients.

To conclude, we tested GSECA under several conditions using distinct simulated and real RNA-seq datasets and different collections of gene sets. Our findings concordantly indicate that GSECA can improve the comprehensive identification of relevant biological processes that are altered in complex phenotypes. In particular, GSECA can detect functionally related and relevant altered cell mechanisms in a condition of interest considering more heterogeneous cohorts as compared to other available methods. By boosting signal-to-noise ratio, GSECA can successfully manage the heterogeneity of thousands of samples and provides useful insights on clinical and biological patterns proper of a phenotype. In this work we introduced the paradigm shift of '*less is more*' in treating large heterogenous RNA-seq datasets and showed that it improves the detection of the altered biological processes in the phenotype of interest.

DATA AVAILABILITY

GSECA is implemented as an R/Shiny application and is freely available on GitHub (https://github.com/matteocereda/GSECA) to be run locally or through a shiny web interface. The EM algorithm for finite Gaussian Mixture is provided by the implemented method in the R package 'mixtools' (95). The *normalmixEM* function was used with default parameters and random initialization.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Matteo Osella, Francesca Ciccarelli and Lara Zandanel for discussions and comments on the manuscript. *Author contributions:* M.Ce. conceived and directed the study; M.Ce. and S.O. supervised the study; M.Ce., A.L. and M.Ca. implemented the method; A.L., S.P., M.D.G., F.P., P.R. and M.Ce. analysed the data; A.L., S.P., M.D.G., P.R., S.O. and M.Ce. wrote the manuscript.

FUNDING

Italian Association for Cancer Research (AIRC) [MFAG 20566 to M.Ce. and IG-20240 to S.O.]. Funding for open access charge: Italian Association for Cancer Research.
Conflict of interest statement. None declared.

REFERENCES

1. Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLos Genet.*, **3**, 1724–1735.
2. Chakravarthy,A., Furness,A., Joshi,K., Ghorani,E., Ford,K., Ward,M.J., King,E.V., Lechner,M., Marafioti,T., Quezada,S.A. *et al.* (2018) Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat. Commun.*, **9**, 3220.
3. May,M. (2017) Big biological impacts from big data. *Science*, **344**, 1298–1301.
4. Rahmatallah,Y., Emmert-Streib,F. and Glazko,G. (2016) Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. *Brief. Bioinform.*, **17**, 393–407.
5. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
6. Cereda,M., Gambardella,G., Benedetti,L., Iannelli,F., Patel,D., Basso,G., Guerra,R.F., Mourikis,T.P., Puccio,I., Sinha,S. *et al.* (2016) Patients with genetically heterogeneous synchronous colorectal cancer carry rare damaging germline mutations in immune-related genes. *Nat. Commun.*, **7**, 12072.
7. McGranahan,N. and Swanton,C. (2017) Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, **168**, 613–628.
8. Cloney,R. (2017) Cancer genomics: single-cell RNA-seq to decipher tumour architecture. *Nat. Rev. Genet.*, **18**, 2–3.
9. Perkel,J.M. (2017) Single-cell sequencing made simple. *Nature*, **547**, 125–126.
10. Turajlic,S., Sottoriva,A., Graham,T. and Swanton,C. (2019) Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.*, **20**, 404–416.
11. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. U.S.A.*, **102**, 15545–15550.
12. Barbie,D.A., Tamayo,P., Boehm,J.S., Kim,S.Y., Moody,S.E., Dunn,I.F., Schinzel,A.C., Sandy,P., Meylan,E., Scholl,C. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.
13. Hänzelmann,S., Castelo,R. and Guinney,J. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.

14. Lee,E., Chuang,H.-Y., Kim,J.-W., Ideker,T. and Lee,D. (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.
15. Tomfohr,J., Lu,J. and Kepler,T.B. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.
16. Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.
17. Tarca,A.L., Bhatti,G. and Romero,R. (2013) A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One*, **8**, e79217.
18. Hebenstreit,D., Fang,M., Gu,M., Charoensawan,V., van Oudenaarden,A. and Teichmann,S.A. (2011) RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.*, **7**, 497–497.
19. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
20. Xiong,Q., Mukherjee,S. and Furey,T.S. (2014) GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data. *Sci. Rep.*, **4**, 6347.
21. Liu,H., Hussain,F., Tan,C.L. and Dash,M. (2002) Discretization: an enabling technique. *Data Mining Knowl. Discov.*, **6**, 393–423.
22. Ramírez-Gallego,S., García,S., Mouriño-Talín,H., Martínez-Rego,D., Bolón-Canedo,V., Alonso-Betanzos,A., Benítez,J.M. and Herrera,F. (2015) Data discretization: taxonomy and big data challenge. *Wires Data Min. Knowl.*, **6**, 5–21.
23. Zhou,L., Pan,S., Wang,J. and Vasilakos,A.V. (2017) Machine learning on big data: opportunities and challenges. *Neurocomputing*, **237**, 350–361.
24. Demichelis,F., Magni,P., Piergiorgi,P., Rubin,M.A. and Bellazzi,R. (2006) A hierarchical Naïve Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays. *BMC Bioinformatics*, **7**, 514.
25. Helman,P., Veroff,R., Atlas,S.R. and Willman,C. (2004) A Bayesian network classification methodology for gene expression data. *J. Comput. Biol.*, **11**, 581–615.
26. McLachlan,G. and Peel,D. (2000) *Mixtures of Factor Analyzers*. John Wiley & Sons, Inc, Hoboken, NJ.
27. Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B (Methodological)*, **39**, 1–38.
28. Littell,R.C. and Folks,J.L. (1973) Asymptotic optimality of Fisher's method of combining independent tests II. *J. Am. Stat. Assoc.*, **68**, 193–194.
29. Cereda,M., Pozzoli,U., Rot,G., Juvan,P., Schweitzer,A., Clark,T. and Ule,J. (2014) RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. *Genome Biol.*, **15**, R20.
30. Gambardella,G., Cereda,M., Benedetti,L. and Ciccarelli,F.D. (2017) MEGA-V: detection of variant gene sets in patient cohorts. *Bioinformatics*, **33**, 1248–1249.
31. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
32. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
33. Schwarz,J.M., Rodelsperger,C., Schuelke,M. and Seelow,D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
34. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
35. Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.
36. Shihab,H.A., Gough,J., Cooper,D.N., Stenson,P.D., Barker,G.L., Edwards,K.J., Day,I.N. and Gaunt,T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57–65.
37. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
38. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
39. Garber,M., Guttman,M., Clamp,M., Zody,M.C., Friedman,N. and Xie,X. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, i54–i62.
40. Liu,X., Wu,C., Li,C. and Boerwinkle,E. (2016) dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.*, **37**, 235–241.
41. Network,C.G.A. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
42. Cereda,M., Sironi,M., Cavalleri,M. and Pozzoli,U. (2011) GeCo++: a C++ library for genomic features computation and annotation in the presence of variants. *Bioinformatics*, **27**, 1313–1315.
43. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
44. Reimand,J., Arak,T., Adler,P., Kolberg,L., Reisberg,S., Peterson,H. and Vilo,J. (2016) g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.*, **44**, W83–W89.
45. Goeman,J.J., van de Geer,S.A., de Kort,F. and van Houwelingen,H.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.
46. Wu,D., Lim,E., Vaillant,F., Asselin-Labat,M.-L., Visvader,J.E. and Smyth,G.K. (2010) ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, **26**, 2176–2182.
47. Geistlinger,L., Csaba,G. and Zimmer,R. (2016) Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. *BMC Bioinformatics*, **17**, 45.
48. Miller,A.J. (1990) Chapman & Hall/CRC Monographs on Statistics and Applied Probability. 2nd Edition. *Subset selection in regression*. Chapman and Hall/CRC, p. 256.
49. Grömping,U. (2006) Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw.*, **17**, 27.
50. Lindeman,R.H., Merenda,P.F. and Gold,R.Z. (1980) *Introduction to Bivariate and Multivariate Analysis*. Scott Foresman, Glenview, IL.
51. Cancer Genome Atlas Research, N. (2015) The molecular taxonomy of primary prostate cancer. *Cell*, **163**, 1011–1025.
52. Thorsson,V., Gibbs,D.L., Brown,S.D., Wolf,D., Bortone,D.S., Ou Yang,T.H., Porta-Pardo,E., Gao,G.F., Plaisier,C.L., Eddy,J.A. *et al.* (2018) The immune landscape of cancer. *Immunity*, **48**, 812–830.
53. Angelova,M., Charoentong,P., Hackl,H., Fischer,M.L., Snajder,R., Krogsdam,A.M., Waldner,M.J., Bindea,G., Mlecnik,B., Galon,J. *et al.* (2015) Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.*, **16**, 64.
54. D'Agostino,R.B., Chase,W. and Belanger,A. (1988) The appropriateness of some common procedures for testing the equality of two independent binomial populations. *Am. Stat.*, **42**, 198–202.
55. Van Rijsbergen,C.J. (1979) *Information Retrieval*. Butterworth-Heinemann.
56. Dimitrova,E.S., Licona,M.P., McGee,J. and Laubenbacher,R. (2010) Discretization of time series data. *J. Comput. Biol.*, **17**, 853–868.
57. Taylor,B.S., Schultz,N., Hieronymus,H., Gopalan,A., Xiao,Y., Carver,B.S., Arora,V.K., Kaushik,P., Cerami,E., Reva,B. *et al.* (2010) Integrative genomic profiling of human prostate cancer. *Cancer Cell*, **18**, 11–22.
58. Robinson,D., Van Allen,E.M., Wu,Y.-M., Schultz,N., Lonigro,R.J., Mosquera,J.M., Montgomery,B., Taplin,M.-E., Pritchard,C.C., Attard,G. *et al.* (2015) Integrative clinical genomics of advanced prostate cancer. *Cell*, **161**, 1215–1228.
59. Yuan,T.L. and Cantley,L.C. (2008) PI3K pathway alterations in cancer: variations on a theme. *Oncogene*, **27**, 5497–5510.
60. Song,M.S., Salmena,L. and Pandolfi,P.P. (2012) The functions and regulation of the PTEN tumour suppressor. *Nat. Rev. Mol. Cell Biol.*, **13**, 283–296.
61. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.

62. Chen,Z.H., Zhu,M., Yang,J., Liang,H., He,J., He,S., Wang,P., Kang,X., McNutt,M.A., Yin,Y. *et al.* (2014) PTEN interacts with histone H1 and controls chromatin condensation. *Cell Rep.*, **8**, 2003–2014.
63. Bononi,A., Bonora,M., Marchi,S., Missiroli,S., Poletti,F., Giorgi,C., Pandolfi,P.P. and Pinton,P. (2013) Identification of PTEN at the ER and MAMs and its regulation of Ca2+ signaling and apoptosis in a protein phosphatase-dependent manner. *Cell Death Differ.*, **20**, 1631–1643.
64. Wang,M.-H., Sun,R., Zhou,X.-M., Zhang,M.-Y., Lu,J.-B., Yang,Y., Zeng,L.-S., Yang,X.-Z., Shi,L., Xiao,R.-W. *et al.* (2018) Epithelial cell adhesion molecule overexpression regulates epithelial-mesenchymal transition, stemness and metastasis of nasopharyngeal carcinoma cells via the PTEN/AKT/mTOR pathway. *Cell Death Dis.*, **9**, 2.
65. Mulholland,D.J., Kobayashi,N., Ruscetti,M., Zhi,A., Tran,L.M., Huang,J., Gleave,M. and Wu,H. (2012) Pten loss and RAS/MAPK activation cooperate to promote EMT and metastasis initiated from prostate cancer stem/progenitor cells. *Cancer Res.*, **72**, 1878–1889.
66. Zhang,Y. and Cheung,Y.-M. (2014) Discretizing numerical attributes in decision tree for big data analysis. *2014 IEEE International Conference on Data Mining Workshop*. IEEE, pp. 1150–1157.
67. Hill,R. and Wu,H. (2009) PTEN, stem cells, and cancer stem cells. *J. Biol. Chem.*, **284**, 11755–11759.
68. Suzuki,A., Kaisho,T., Ohishi,M., Tsukio-Yamaguchi,M., Tsubata,T., Koni,P.A., Sasaki,T., Mak,T.W. and Nakano,T. (2003) Critical roles of Pten in B cell homeostasis and immunoglobulin class switch recombination. *J. Exp. Med.*, **197**, 657–667.
69. Newton,R.H. and Turka,L.A. (2012) Regulation of T cell homeostasis and responses by pten. *Front. Immunol.*, **3**, 151.
70. Cao,X., Wei,G., Fang,H., Guo,J., Weinstein,M., Marsh,C.B., Ostrowski,M.C. and Tridandapani,S. (2004) The inositol 3-phosphatase PTEN negatively regulates Fc gamma receptor signaling, but supports Toll-like receptor 4 signaling in murine peritoneal macrophages. *J. Immunol.*, **172**, 4851–4857.
71. Garg,R., Blando,J.M., Perez,C.J., Abba,M.C., Benavides,F. and Kazanietz,M.G. (2017) Protein Kinase C epsilon cooperates with PTEN loss for prostate tumorigenesis through the CXCL13-CXCR5 Pathway. *Cell Rep.*, **19**, 375–388.
72. Ortega-Molina,A. and Serrano,M. (2013) PTEN in cancer, metabolism, and aging. *Trends Endocrinol. Metab.*, **24**, 184–189.
73. Crackower,M.A., Oudit,G.Y., Kozieradzki,I., Sarao,R., Sun,H., Sasaki,T., Hirsch,E., Suzuki,A., Shioi,T., Irie-Sasaki,J. *et al.* (2002) Regulation of myocardial contractility and cell size by distinct PI3K-PTEN signaling pathways. *Cell*, **110**, 737–749.
74. Soundararajan,R., Pearce,D. and Ziera,T. (2012) The role of the ENaC-regulatory complex in aldosterone-mediated sodium transport. *Mol. Cell Endocrinol.*, **350**, 242–247.
75. Milella,M., Falcone,I., Conciatori,F., Cesta Incani,U., Del Curatolo,A., Inzerilli,N., Nuzzo,C.M.A., Vaccaro,V., Vari,S., Cognetti,F. *et al.* (2015) PTEN: Multiple Functions in Human Malignant Tumors. *Front. Oncol.*, **5**, 24.
76. Westin,S.N., Ju,Z., Broaddus,R.R., Krakstad,C., Li,J., Pal,N., Lu,K.H., Coleman,R.L., Hennessy,B.T., Klempner,S.J. *et al.* (2015) PTEN loss is a context-dependent outcome determinant in obese and non-obese endometrioid endometrial cancer patients. *Mol. Oncol.*, **9**, 1694–1703.
77. Cancer Genome Atlas Research, N. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
78. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
79. Chen,L. and Guo,D. (2017) The functions of tumor suppressor PTEN in innate and adaptive immunity. *Cell Mol. Immunol.*, **14**, 581–589.
80. Jamaspishvili,T., Berman,D.M., Ross,A.E., Scher,H.I., De Marzo,A.M., Squire,J.A. and Lotan,T.L. (2018) Clinical implications of PTEN loss in prostate cancer. *Nat. Rev. Urol.*, **15**, 222–234.
81. Zhao,J., Chen,A.X., Gartrell,R.D., Silverman,A.M., Aparicio,L., Chu,T., Bordbar,D., Shan,D., Samanamud,J., Mahajan,A. *et al.* (2019) Immune and genomic correlates of response to anti-PD-1 immunotherapy in glioblastoma. *Nat. Med.*, **25**, 462–469.
82. Wolf,D.M., Lenburg,M.E., Yau,C., Boudreau,A. and van 't Veer,L.J. (2014) Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. *PLoS One*, **9**, e88309.
83. Lausen,B. and M.,S. (1992) Maximally selected rank statistics. *Biometrics*, **48**, 73–85.
84. Toso,A., Revandkar,A., Di Mitri,D., Guccini,I., Proietti,M., Sarti,M., Pinton,S., Zhang,J., Kalathur,M., Civenni,G. *et al.* (2014) Enhancing chemotherapy efficacy in Pten-deficient prostate tumors by activating the senescence-associated antitumor immunity. *Cell Rep.*, **9**, 75–89.
85. Armstrong,C.W.D., Maxwell,P.J., Ong,C.W., Redmond,K.M., McCann,C., Neisen,J., Ward,G.A., Chessari,G., Johnson,C., Crawford,N.T. *et al.* (2016) PTEN deficiency promotes macrophage infiltration and hypersensitivity of prostate cancer to IAP antagonist/radiation combination therapy. *Oncotarget*, **7**, 7885–7898.
86. Leavy,O. (2015) Regulatory T cells. The PTEN stabilizer. *Nat. Rev. Immunol.*, **15**, 71–71.
87. George,S., Miao,D., Demetri,G.D., Adeegbe,D., Rodig,S.J., Shukla,S., Lipschitz,M., Amin-Mansour,A., Raut,C.P., Carter,S.L. *et al.* (2017) Loss of PTEN is associated with resistance to Anti-PD-1 checkpoint blockade therapy in metastatic uterine leiomyosarcoma. *Immunity*, **46**, 197–204.
88. Peng,W., Chen,J.Q., Liu,C., Malu,S., Creasy,C., Tetzlaff,M.T., Xu,C., McKenzie,J.A., Zhang,C., Liang,X. *et al.* (2016) Loss of PTEN promotes resistance to T Cell-Mediated immunotherapy. *Cancer Discov.*, **6**, 202–216.
89. Tilot,A.K., Bebek,G., Niazi,F., Altemus,J.B., Romigh,T., Frazier,T.W. and Eng,C. (2016) Neural transcriptome of constitutional Pten dysfunction in mice and its relevance to human idiopathic autism spectrum disorder. *Mol. Psychiatry*, **21**, 118–125.
90. Fan,J., Han,F. and Liu,H. (2014) Challenges of Big Data analysis. *Natl. Sci. Rev.*, **1**, 293–314.
91. Marron,J.S. (2017) Big Data in context and robustness against heterogeneity. *Econo. Stat.*, **2**, 73–80.
92. Tamayo,P., Steinhardt,G., Liberzon,A. and Mesirov,J.P. (2016) The limitations of simple gene set enrichment analysis assuming gene independence. *Stat. Methods Med. Res.*, **25**, 472–487.
93. Zhan,T., Rindtorff,N. and Boutros,M. (2017) Wnt signaling in cancer. *Oncogene*, **36**, 1461–1473.
94. Sharma,P., Hu-Lieskovan,S., Wargo,J.A. and Ribas,A. (2017) Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell*, **168**, 707–723.
95. Benaglia,T., Chauveau,D., Hunter,D. and Young,D. (2009) mixtools: an R package for analyzing finite mixture models. *J. Stat. Softw.*, **32**, 1–29.