Published OnlineFirst September 19, 2018; DOI: 10.1158/0008-5472.CAN-18-1669

Convergence and Technologies: *Resource Report*

Cancer Research

CANCERTOOL: A Visualization and Representation Interface to Exploit Cancer Datasets

Ana R. Cortazar[1,2], Veronica Torrano[1,2], Natalia Martín-Martín[1,2], Alfredo Caro-Maldonado[1], Laura Camacho[1,3], Ivana Hermanova[1], Elizabeth Guruceaga[2,4], Luis F. Lorenzo-Martín[2,5,6], Ruben Caloto[2,5,6], Roger R. Gomis[2,7,8], Iñigo Apaolaza[9], Victor Quesada[2,10], Jan Trka[11,12], Antonio Gomez-Muñoz[3], Silvestre Vincent[2,13,14,15], Xose R. Bustelo[2,5,6], Francisco J. Planes[9], Ana M. Aransay[1,16], and Arkaitz Carracedo[1,2,3,17]

Abstract

With the advent of OMICs technologies, both individual research groups and consortia have spear-headed the characterization of human samples of multiple pathophysiologic origins, resulting in thousands of archived genomes and transcriptomes. Although a variety of web tools are now available to extract information from OMICs data, their utility has been limited by the capacity of nonbioinformatician researchers to exploit the information. To address this problem, we have developed CANCERTOOL, a web-based interface that aims to overcome the major limitations of public transcriptomics dataset analysis for highly prevalent types of cancer (breast, prostate, lung, and colorectal). CANCERTOOL provides rapid and comprehensive visualization of gene expression data for the gene(s) of interest in well-annotated cancer datasets. This visualization is accompanied by generation of reports customized to the interest of the researcher (e.g., editable figures, detailed statistical analyses, and access to raw data for reanalysis). It also carries out gene-to-gene correlations in multiple datasets at the same time or using preset patient groups. Finally, this new tool solves the time-consuming task of performing functional enrichment analysis with gene sets of interest using up to 11 different databases at the same time. Collectively, CANCERTOOL represents a simple and freely accessible interface to interrogate well-annotated datasets and obtain publishable representations that can contribute to refinement and guidance of cancer-related investigations at all levels of hypotheses and design.

Significance: In order to facilitate access of research groups without bioinformatics support to public transcriptomics data, we have developed a free online tool with an easy-to-use interface that allows researchers to obtain quality information in a readily publishable format. *Cancer Res; 78(21); 6320–8. ©2018 AACR.*

Introduction

Cancer encompasses a large collection of diseases that extensively vary in terms of mutation load, driver pathobiologic programs, metabolic needs, and microenvironmental constraints (1). This heterogeneity is largely responsible for the current challenges we face in terms of patient classification and effective treatments (2). The mutational backpack of tumors has been the main focus of research in recent years (3). However, current data indicate that understanding of genome-wide transcriptional programs can also provide important information for patient stratification, diagnosis, and the determination of possible therapeutic protocols. In this context, normalized procedures have been established to make transcriptomic data publicly available (4). However, the utilization of these databases to extract information still represents an important limitation for research groups that do not have adjacent bioinformatics support. To bypass this problem, various

[1]CIC bioGUNE, Bizkaia Technology Park, Bizkaia, Spain. [2]CIBERONC, Madrid, Spain. [3]Biochemistry and Molecular Biology Department, University of the Basque Country (UPV/EHU), Bilbao, Spain. [4]Bioinformatics Unit, Center for Applied Medical Research, University of Navarra, Pamplona, Spain. [5]Centro de Investigación del Cáncer, Consejo Superior de Investigaciones Científicas (CSIC), University of Salamanca, Salamanca, Spain. [6]Instituto de Biología Molecular y Celular del Cáncer, Consejo Superior de Investigaciones Científicas (CSIC), University of Salamanca, Salamanca, Spain. [7]Oncology Programme, Institute for Research in Biomedicine (IRB-Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [8]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [9]University of Navarra, Tecnun School of Engineering, San Sebastián, Spain. [10]Departamento de Bioquímica y Biología Molecular, Universidad de Oviedo, Oviedo, Spain. [11]CLIP-Childhood Leukaemia Investigation Prague and Second Faculty of Medicine, Charles University, Prague, Czech Republic. [12]University Hospital Motol, Prague, Czech Republic. [13]University of Navarra, Department of Histology and Pathology, Pamplona, Spain. [14]IdiSNA, Navarra Institute for Health Research, Pamplona, Spain. [15]Center for Applied Medical Research, Program of Solid Tumors, University of Navarra, Pamplona, Spain. [16]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Madrid, Spain. [17]Ikerbasque, Basque Foundation for Science, Bilbao, Spain.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

A.M. Aransay and A.Carracedo contributed equally to this article.

The codebase and databases are freely available upon request (http://web.bioinformatics.cicbiogune.es/CANCERTOOL/).

Corresponding Author: Arkaitz Carracedo, CIC bioGUNE, Derio, Vizcaya 48160, Spain. Phone: 349-4657-2517; Fax: 349-4406-1301; E-mail: acarracedo@cicbiogune.es

doi: 10.1158/0008-5472.CAN-18-1669

Published OnlineFirst September 19, 2018; DOI: 10.1158/0008-5472.CAN-18-1669

computational tools and portals have been created. For example, the GEO database (https://www.ncbi.nlm.nih.gov/geo/; ref. 5) has been developed by the NCBI to facilitate the public access to functional genomics data (including gene expression data). The tools available in this database allow the simple visualization of data upon specific queries. However, specialized personnel are still required to extract information on the cancer type of interest, export the data obtained, perform clinical associations, or to obtain publication-grade representations from the data generated. Some of the latter aspects are fulfilled by Oncomine (6), a commercial platform that offers several tools for analyzing gene expression in multiple sets of data from independent studies at the same time. It also allows searching information through various filters (such as the type of cancer, sample types, names of specific genes, etc.) and returns results from multiple analyses carried out using gene information according to the requirements set forth by the user. A main problem of this tool is that it requires a costly subscription to get access to all its capabilities. It is also far from being user friendly for nonspecialists. More recently, cBioPortal has become the most attractive portal for cancer OMICs analysis (7, 8). This tool is free and enables the user to browse through multiple datasets and query multiple genes. It also provides a user-friendly representation for data interpretation. However, reanalyses from raw data are frequently required for publication-quality figures and the browsing for information regarding gene expression alterations in cancer is still time consuming owing to the multiple options and datasets available. Whereas non-bioinformatician cancer researchers could access all required transcriptomics information for major cancer types with the aforementioned tools, the process still requires considerable training, over dozens of clicks, and additional time for preparing publication-quality figures.

In this article, we report CANCERTOOL (http://web.bioinformatics.cicbiogune.es/CANCERTOOL), a new portal that aims at solving the foregoing problems. This tool focuses on the four major tumor types (breast, prostate, colorectal, and lung) in its first version, although it is engineered to quickly incorporate new studies from either the aforementioned tumors or other cancer types. Importantly, the datasets contained in CANCERTOOL have been carefully curated to offer various types of clinical data related to each cancer such as disease progression, pathologic, and molecular annotations. All these results are presented in a format that allows the user to screen through tens of candidate genes within minutes, as well as to perform customized analyses for retrieval of high-quality representations, detailed statistics, and access to raw data for reanalysis of the selected hits. CANCERTOOL offers the opportunity of performing comparative gene expression analyses among investigator-selected conditions (e.g., sample type, disease stage, and pathologic and molecular features) and to estimate the association of candidate gene expression to disease progression. To provide a complete toolbox in a stereotypic OMICs cancer research project, CANCERTOOL includes a functional enrichment package with access to 11 databases that can be exploited with either in-house experimental- or CANCERTOOL-derived gene sets. As such, CANCERTOOL aims at providing access to transcriptomics cancer data selected for rich clinical annotation not currently offered by the tools discussed above. Ultimately, this tool focuses on free, rapid, and comprehensive visualization analyses of transcriptomics results that are ready to use for cancer researchers that lack bioinformatics support.

Materials and Methods

Available datasets and its normalization

For the development of CANCERTOOL, data from the public repository GEO (5), cBioPortal (for METABRIC, cbioportal.org), and from the project The Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov/; ref. 4) were obtained. Specifically, gene expression and phenotypic data from patients with breast, colorectal, lung, or prostate cancer were retrieved. The datasets' identification and related cancer information can be found in Table 1 and in the Datasets section in CANCERTOOL. In addition, the Datasets section also offers full access to all phenotypic information included in every dataset for all patients. The raw gene expression data available in the repositories, in the form of fluorescence intensity or number of sequencing reads, were first $\log_2$ transformed and quartile normalized (9).

Data coming from the GEO repository were downloaded as series matrix, and were $\log_2$ transformed and quartile normalized when needed. METABRIC data were downloaded from cBioPortal as $\log_2$ transformed and normalized, while TCGA data were downloaded as upper quartile normalized RSEM count, which had been $\log_2$ transformed. Regarding data donated by collaborators, they have been treated as mentioned in Vallejo and colleagues (10).

CANCERTOOL architecture

mRNA and long intergenic noncoding RNA (LINC) expression levels and clinical data were indexed and queried via MySQL relational database with a MyISAM engine, improving the speed for the retrieval of results. CANCERTOOL is located on a Linux/Apache server to enable both enhanced stability and security. Its architecture is built around a three-tier model: the presentation, the logic, and the database tiers. The presentation tier, implemented in PHP/CSS and JavaScript, handles user's requests and

Table 1. All the available cancer types and datasets considered in CANCERTOOL

Breast cancer	Colorectal cancer	Lung adenocarcinoma	Prostate cancer
Ivshina et al. (20)	Colonomics https://www.colonomics.org/	Chitale et al. (21)	Glinksy et al. (22)
Lu et al. (23)	Jorissen et al. (24)	Okayama et al. (25)	Grasso et al. (26)
METABRIC (27)	Kemper et al. (28)	Shedden et al. (29)	Lapointe et al. (30)
Pawitan et al. (31)	Laibe et al. (32)	TCGA (RNA-seq) https://cancergenome.nih.gov/	Taylor et al. (33)
TCGA (microarray) https://cancergenome.nih.gov/	Marisa et al. (34)	Wilkerson et al. (35)	TCGA (RNA-seq) https://cancergenome.nih.gov/
Wang et al. (36)	Roepman et al. (37)		Tomlins et al. (38)
	TCGA (RNA-seq) https://cancergenome.nih.gov/		Varambally et al. (39)

Published OnlineFirst September 19, 2018; DOI: 10.1158/0008-5472.CAN-18-1669

displays data results. The logic tier contains most of the functions to handle data transfer between the presentation and the database tier, being implemented using Perl and R scripting languages. The database tier is where data are located and is responsible for accessing them.

Statistical analysis

CANCERTOOL includes various statistical calculations to compare the levels of expression of a gene among different types of patients, to study the disease-free survival depending on the expression levels of target genes, to make correlations between pairs of genes, and to perform enrichment analysis of gene sets.

For the comparison of mRNA and/or LINC expression levels among different types of patient specimens, normality was assessed and parametric tests were set to compare means among groups. CANCERTOOL performs Student *t* test to interrogate the differences between two groups. For comparisons among means of more than two groups of specimens for a given gene, ANOVA test is performed. In custom studies, a *post hoc* analysis is included using Bonferroni and Tukey HSD (11, 12). In addition, when data from more than two datasets are produced in a given analysis, Edgington method is applied (the sum of *P* values; ref. 13) to ascertain whether the integration of all datasets yield a significant difference. Also in custom analyses, *P* values have been adjusted using Benjamini–Hochberg method. In all the cases, $P \leq 0.05$ is considered as significant. All the analyses were performed with R, using stats (14) package for adjusting and *post hoc* analyses, metap (https://rdrr.io/cran/metap/) package for the calculation of Edgington method, and ggplot2 for the violin plots (15).

CANCERTOOL provides to distinct statistical analyses for the Correlations section based on the criteria of the end-user. Pearson and/or the Spearman correlation coefficient is calculated for every pair of genes, and the statistical significance is provided. For correlations that generate results in more than two datasets, a "Coherence" value is also calculated. This parameter informs the user about the consistent directional correlation present in the analysis when integrating the results from all available datasets. An accountable correlation is estimated when the $P \leq 0.05$, and the correlation coefficient is greater than 0.2 (for direct correlations) or lower than −0.2 (inverse correlations). Analyses that present the accountable correlations with the same directionality in more than 50% of the available datasets are flagged as "Coherent". In custom analyses, *P* values have been adjusted using FDR method. All the correlation analyses and graphs were performed with R while correlation heatmaps were drawn using a custom version of the heatmap.2 function from package gplots (https://cran.r-project.org/web/packages/gplots/index.html).

For survival analyses, CANCERTOOL performs a quartile-based separation of patients on the basis of the expression of the gene of interest. Next, the survival curves are represented using the Kaplan–Meier estimator as quartile 1 (Q1, 25% of patient specimens with the lowest expression), quartile 4 (Q4, 25% of patient specimens with the highest expression), and Q2+Q3 (the 50% of patient specimens with expression ranging between Q1 and Q4). The statistical significance is provided by the Mantel–Cox (also known as log-rank) test. This test was selected because it assumes the randomness of the possible censorship (16). All the survival analyses and graphs were performed with R (https://cran.r-project.org/web/packages/survival/citation.html) and $P \leq 0.05$ was considered statistically significant.

Gene enrichment analysis employs a hypergeometric test and FDR method to adjust *P* values. This methodology was chosen because data come from a simple random sampling without replacement. Please note that whereas previous analyses are associated to datasets contained in CANCERTOOL, gene enrichment can be performed using CANCERTOOL-derived information (obtained in the complementary tools) or with user-defined gene sets. Gene enrichment analysis can be performed in the databases indicated in Supplementary Table S1. An adjusted *P* (P_{adj}) value ≤ 0.05 was considered statistically significant. All the calculations have been performed with R.

Results

CANCERTOOL is a freely accessible web tool, which allows the user to query a database formed by manually curated transcriptomics cancer datasets for the most prevalent tumor types. The tool is organized in three different sections: Basic Analyses, Correlations, and Gene Enrichment. It also allows the user to access the Manual, Datasets, information related to the developers, and contact information in the Datasets (Help, About Us, Contact Us and Citation sections).

Basic analyses

This section has been designed to satisfy an important requirement for cancer researchers, namely browsing rapidly through the results in the most prevalent tumor types. A summary PDF of the results is provided with the aim of complying with a rapid *Go/No-Go* decision–making strategy. The current output of CANCERTOOL is compatible with the selection of best candidates in the tumor type of interest based on the mRNA and/or LINC expression profile among dozens of queried genes in a matter of minutes. The end-user receives visual representations related to: (i) basic statistical analyses for the gene(s) of interest; (ii) comparisons of the relative expression of the gene(s) under analysis in tumor versus healthy tissue, in different pathologic features (e.g. stage, Gleason score, and location), molecular characteristics of the tumors (e.g. ER status, *KRAS*, and *EGFR* mutation status), molecular subtype (e.g. luminal, basal, and $HER2^+$ in the case of breast tumor samples), and disease progression (i.e., primary tumor vs. metastasis, disease-free/metastasis-free/overall survival). The mRNA and/or LINC expression comparisons among groups of specimens are provided as easily interpretable violin plots that, in each case, provide additional information on sample size estimation. Another advantage is that it avoids the limitations usually observed when using other type of representations (e.g., dot plots) when the datasets encompass large sample size (17). This representation is visually appealing and informative, and it is a distinctive value when compared with other visualization tools (Supplementary Fig. S1). Survival analysis is provided using a Kaplan–Meier estimator that divides the sample set according to the expression of the mRNA and/or LINC of interest in quartiles. With this output, CANCERTOOL facilitates rapid decision-making by the user without any type of previous knowledge on bioinformatics. Furthermore, it also enables further analyses of the mRNA and/or LINC(s) that comply with preestablished selection criteria according to the needs of the customer. In Fig. 1, we depict an illuminating example of the analysis of a gene identified as potential regulator of prostate cancer biology using this type of Basic Analyses tool. This gene was subsequently corroborated as an important player in this tumor type (18). The

Published OnlineFirst September 19, 2018; DOI: 10.1158/0008-5472.CAN-18-1669

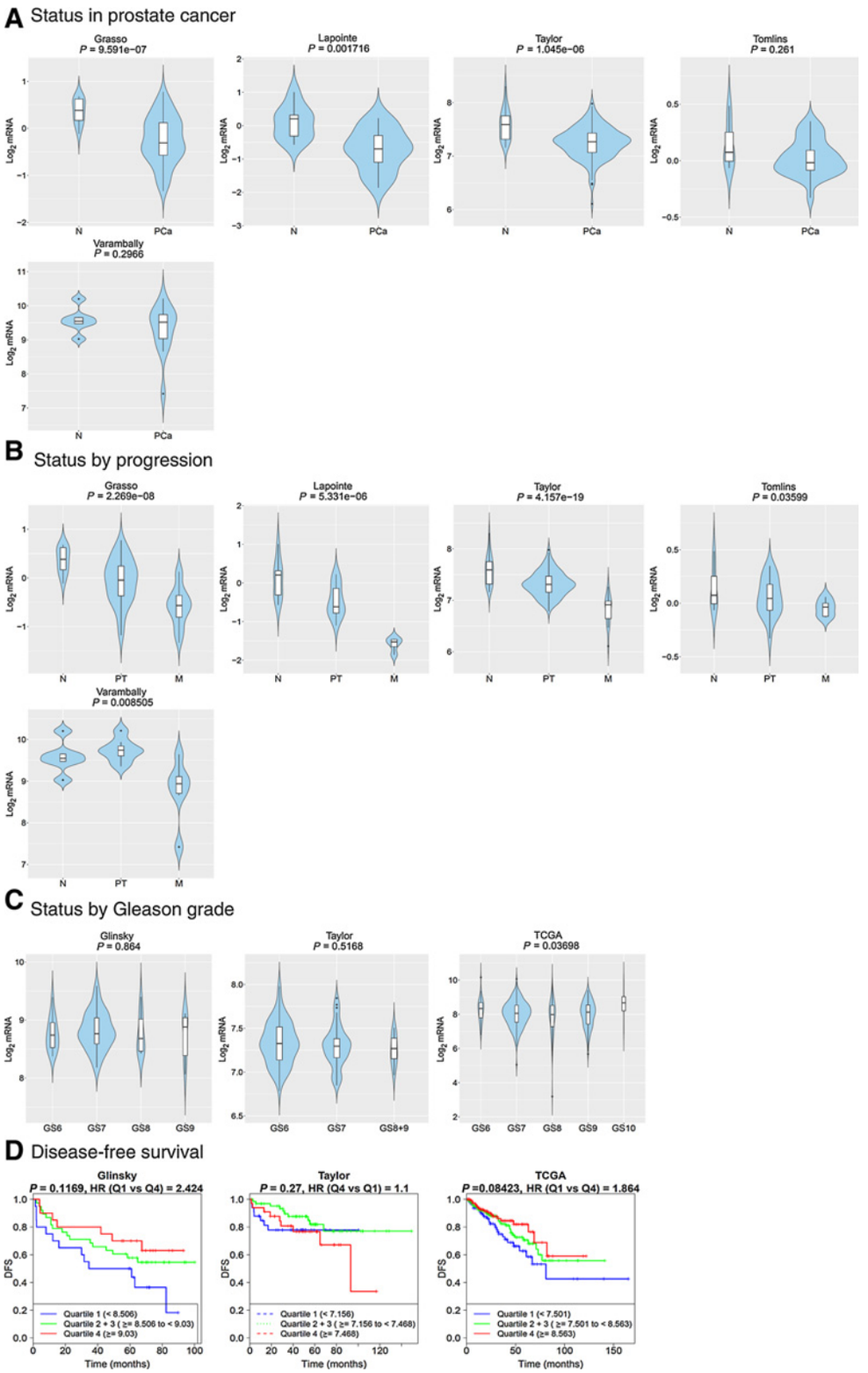

Figure 1.

A representative summary output of the Basic Analysis section in CANCERTOOL for the gene *MITF* in prostate cancer datasets. Violin plots depicting the expression of the gene of interest between nontumoral (N, normal) and prostate cancer (PCa) specimens (**A**), among nontumoral (normal), primary tumor and metastatic (prostate cancer) specimens (**B**), and among prostate cancer specimens of the indicated Gleason grade (GS, Gleason score; **C**) in the indicated datasets. The *y*-axis represents the $\log_2$-normalized gene expression. **D,** Kaplan–Meier curves representing the disease-free survival (DFS) of patient groups selected according to the quartile expression of the gene of interest. Statistical analysis: Student *t* test (**A**), ANOVA (**B** and **C**), and Mantel–Cox test (**D**).

Published OnlineFirst September 19, 2018; DOI: 10.1158/0008-5472.CAN-18-1669

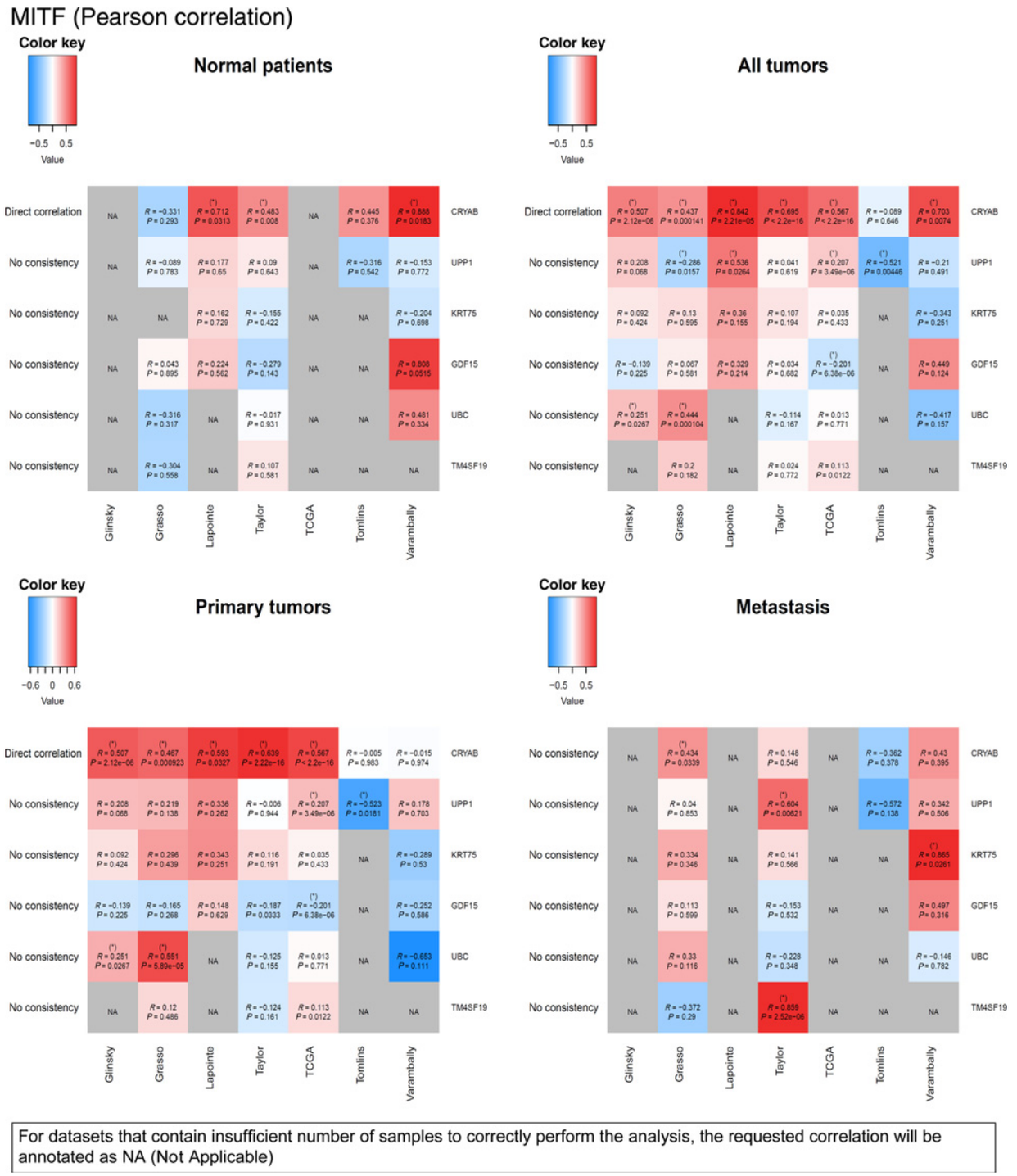

Figure 2.

Representative heatmap provided by CANCERTOOL for the correlation analyses of *MITF* against the indicated genes in multiple prostate cancer datasets. The color code indicates the correlation status between the indicated gene pairs, being red toward 1 and blue toward −1. In case of not applicable (NA), the cell is gray and with no data. Correlations are indicated with asterisks when they comply with the following significance criteria: $P \leq 0.05$ and correlation coefficient greater than 0.2 for direct and lower than −0.2 for inverse correlations. On the left side, the coherence among datasets is shown for each pair of genes (directional correlation in more than 50% of datasets, being $P \leq 0.05$ and correlation coefficient greater than 0.2 for direct and lower than −0.2 for inverse correlations).

Published OnlineFirst September 19, 2018; DOI: 10.1158/0008-5472.CAN-18-1669

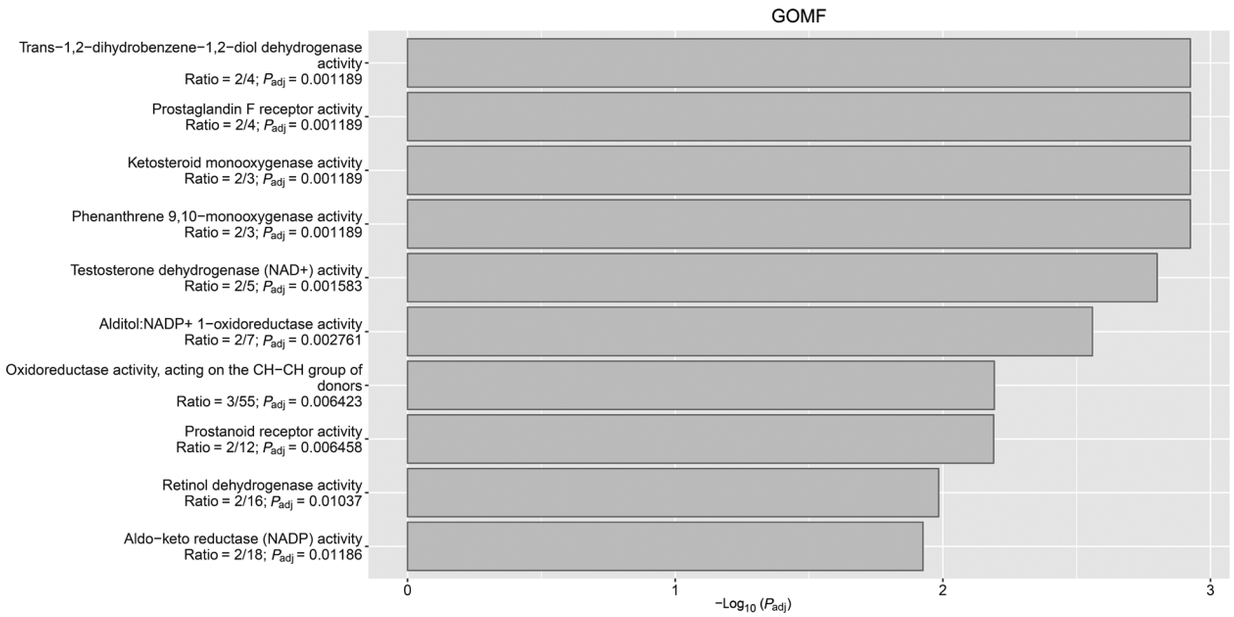

Figure 3.

Example of the output figure provided by the gene enrichment analysis. The results show a histogram from the Molecular Function database within Gene Ontology. The 10 terms with the highest significance in adjusted FDR *P* value, ordered by this field, are included. The *x*-axis indicates the $-\log_{10}(P_{adj}$ value). The *y*-axis includes information relative to the category, the ratio, and the P_{adj} value.

Summary file provides rapid and visual information regarding the downregulation of the Microphthalmia-associated transcription factor (*MITF*) in three out of five prostate cancer datasets analyzed (Fig. 1A). Furthermore, its shows that such a deregulation is associated with the progression of the disease towards metastasis in four out the five datasets analyzed using CANCERTOOL (Fig. 1B). These analyses also make apparent that the expression of this gene does not have any statistically significant correlation with the Gleason score (Fig. 1C). It does predict disease-free survival of prostate cancer when patients are stratified according to the first quartile (Q1) of *MITF* expression (Fig. 1D). However, a conclusive result would require further customized analysis. Importantly, the user can obtain publication-quality images for each of the foregoing study (Fig. 1). It is also possible to conduct more detailed statistical analyses and representations stemming from the raw data in the Custom section. Thus, together with the quartile-based analysis, this custom study allowed us to obtain representations of two groups of *MITF* expression divided by the mean expression of this gene in the cohort of interest (Supplementary Fig. S2). Additional personalized cutoffs can be established through the access to raw data files provided in the custom analysis. Indeed, the inverse association of *MITF* to disease-free survival was significant in two out of three datasets analyzed when the patients within the first quartile of gene expression were compared with the rest of the cohort (18). Of note, in support of the possibility of monitoring the expression of LINCs in CANCERTOOL, Supplementary Fig. S3 depicts the expression of LINC00116 in a dataset per tumor type.

Correlations

CANCERTOOL can calculate and plot gene-to-gene correlations in the annotated tumor types and datasets. The output of this tool has been designed for rapid *Go*/*No-Go* decisions. The tool allows up to 5 (list 1) $\times$ 10 (list 2) gene comparison. From the summary analysis, the user obtains two types of output files: (i) a PDF with the correlation results visualized as a heatmap representation per gene (from list 1) that is subject to correlations against a gene set (list 2), in which significant correlations (that reach a criteria of $P \leq 0.05$ and $-0.2 \geq R \geq 0.2$) are indicated with an asterisk (Fig. 2); (ii) a PDF depicting, for each gene-to-gene correlation, the study in various patient subgroups (e.g. nontumoral tissue, cancer specimens, cancer subtypes, and progression stages; Supplementary Fig. S4). The correlation analyses can be also customized to select either the datasets or the patient subsets where to perform the correlations and the type of statistics (Pearson and/or Spearman). The results are finally presented in several outputs, including high-resolution figures, raw data for reanalysis, and tables including the calculated correlation coefficients and *P* values (and Benjamini–Hochberg correction) for each of the primary datasets utilized by the tool. The table also includes a "coherence" calculation that estimates the robustness of gene-to-gene association across all the datasets that have been interrogated. Coherence estimates consistent correlations when more than 50% of the datasets show a significant and unidirectional correlation with correlation coefficient greater than 0.2 (for direct correlations) or lower than -0.2 (for inverse correlations; Supplementary Table S2). The Correlations section in CANCERTOOL can aid the researcher in the screening and selection of the best potential gene associations to uncover functional implications in cancer (18).

Gene enrichment

Cancer research studies often exploit OMICs-derived data to decipher the molecular mechanisms associated with changes in the expression of either a gene or pathway of interest. The

Published OnlineFirst September 19, 2018; DOI: 10.1158/0008-5472.CAN-18-1669

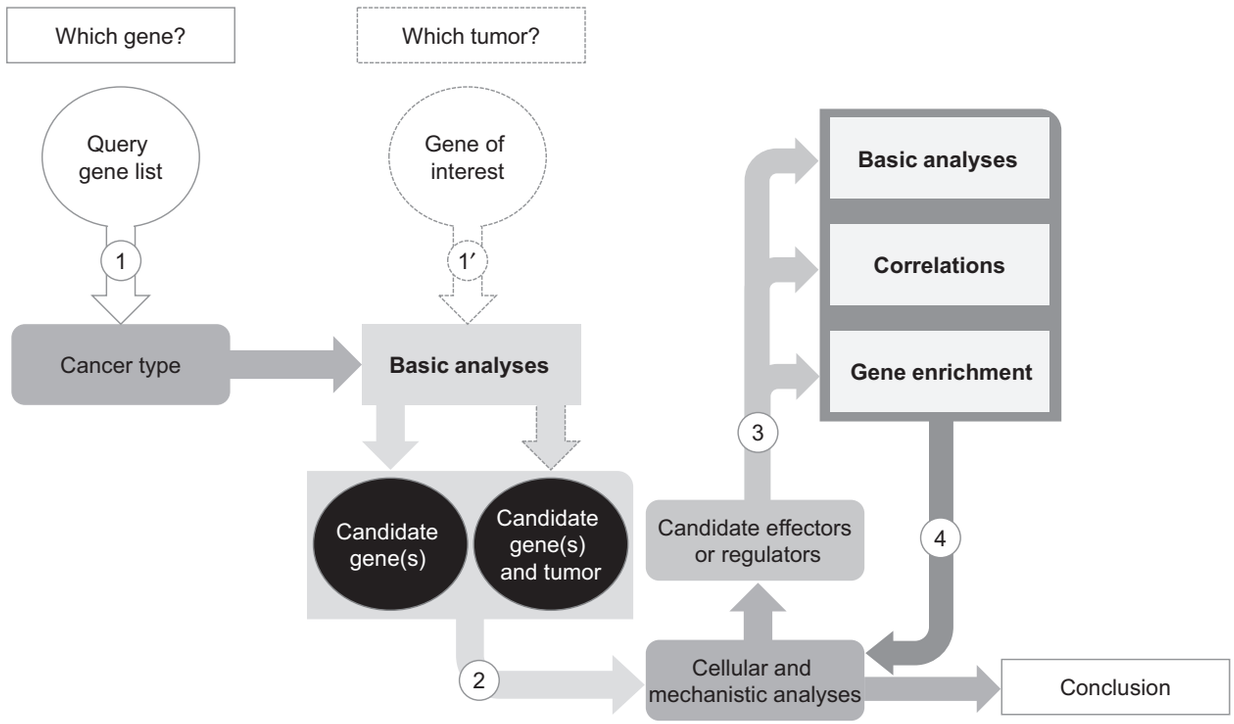

Figure 4.

Example of a workflow integrating CANCERTOOL with empirical studies in cancer research. Step 1/1′, initial selection of genes to be queried in CANCERTOOL. Genes can be studied in a single cancer type (to be selected in the interface) if it is predetermined (step 1), or sequential analysis can be performed to gather information about all cancer types available (step 1′). The results will aid the user in the identification of the best candidate genes for further analysis and the most promising cancer type where the research question is to be developed. Step 2, The selection of candidate genes can be followed by experimental approaches to shed light on the mechanism of action, often accompanied with OMICs strategies to provide a comprehensive view of the molecular alterations associated with the candidate gene(s). Step 3, Once the researcher reaches the identification of potential effectors, upstream regulators or gene lists perturbed upon manipulation of our gene(s) of interest, CANCERTOOL can aid in the identification of most interesting candidates (based on gene expression alterations in patients with cancer, correlations analyses with the gene(s) driving the research project, or enrichment analysis) to identify molecular/pathologic process and common regulatory cues predominant in the queried gene list. Step 4, these analyses can enable the user to further refine the mechanistic hypotheses and to reach relevant conclusions.

result of such analysis usually includes large lists of perturbed genes, transcripts, and/or proteins. Whereas gene-by-gene annotation can be useful to define individual candidate genes at the core of a given molecular mechanism, bioinformatics also offers the possibility of using more integrative analyses from gene lists to unveil pathways and/or regulatory hubs that would otherwise remain hidden. There are several databases that perform complementary enrichments. However, they usually require complex and lengthy analytic steps that are not usually easy to implement by nonspecialists. To tackle this customer demand, CANCERTOOL includes an Additional Analyses section where we offer researchers a simplified and comprehensive access to additional aspects related to gene regulation and functional integration. These tools enable the researcher to round up OMICs-related cancer studies and, following the overall philosophy of all the tools associated with this platform, generate output data in publication-quality images and with the potential of subsequent customer-driven reanalyses. To this end, CANCERTOOL harbors 11 independent enrichment databases, including the basic Gene Ontology analysis [GO; biological process (GOBP), molecular function (GOMF), and cell compartment (GOCC)], pathways and pathophysiologic processes (KEGG, Biocarta, Reactome, Biocarta, Onco, DOSE, HIPC, and Connectivity Map), and the upstream regulatory cue prediction tool (TFT, MIR). Results are presented as a platform-generated spreadsheet per enrichment database that highlights type of main enriched functions, the prevalence of such functions within the gene list, and statistical significance of the associations sieved according to the Benjamini–Hochberg correction (P_{adj} value). The output spreadsheets are further complemented with a visual representation that depicts the 10 most significant functions per database sorted according to P_{adj} value (Fig. 3).

Discussion

The exploitation of data from publicly available datasets has become the bottleneck for researchers that do not have a strong bioinformatics background or facilities. Various tools do offer data browsing and analysis. However, the interpretation and representation of these data is still complex and cumbersome (see Introduction). This issue is particularly important in the context of cancer research, where the availability of data is overwhelming and grows exponentially every year. In this context, the bioinformatics capability of a lab is a differentiating factor to increase the efficacy, the productivity, and the biomedical impact

Published OnlineFirst September 19, 2018; DOI: 10.1158/0008-5472.CAN-18-1669

of the results. Public dataset exploitation can be instrumental for the refinement of hypotheses, the selection of candidate genes (thus reducing the time and cost of exploratory experiments), and the validation of mechanistic studies.

CANCERTOOL has been specifically designed to fulfill this need in cancer research and to complement the capabilities of other existing tools and portals. Its main features include: (i) the ability to provide to users a rapid evaluation of gene expression data for the queried gene(s); (ii) ensuring full availability of the raw data used in the analyses; (iii) providing high-quality and further editable figures; and (iv) making possible functional annotations to facilitate the association of the gene(s) of interest with specific functional networks, hubs, and pathophysiologic programs (Fig. 4). These features are given in an easy-to-use platform that will enable cancer researchers to go from a candidate gene list to the final publication-grade representation of their best candidates in a short time frame. Furthermore, its ability to provide basic gene expression comparisons among patient subgroups and correlations among genes of interest in various datasets enables cancer researchers to maximize the invested time and effort, and in turn to make significant advances in the postulated research question. CANCERTOOL includes clinical and molecular features based on the availability of the datasets of origin. Thus, as more clinical data–rich studies are produced, the parameters of analysis included in this tool will progressively increase.

This tool has been built using public transcriptomics datasets from four major tumor types. This strategy stems from the importance of carefully selecting and curating datasets that are rich in clinical, pathologic, and molecular data associated with each sample. We have prioritized the inclusion of a selected number of datasets to ensure that the interpretation of the results is rapid and efficient without compromising the robustness or translational potential of the results. However, CANCERTOOL has been designed to allow its subsequent expansion according to the needs of cancer researchers. Hence, the pipeline for dataset inclusion in CANCERTOOL is ready to progressively incorporate additional datasets and cancer types. Further improvements can be done at the level of the statistics used and the type of graphical outputs to enrich the summary and custom analyses, always keeping in mind that they must be generated in quick and easy manner (the *Go/No-Go* strategy) to be understood and evaluated. Moreover, the support for gene signatures (average signal of various genes within a functional group) in the sections Basic Analyses and correlations will expand the use of the interface. A similar strategy to the transcriptomics analysis presented herein can be applied to other OMICs layers. As an example, the inclusion of methylation studies could be an invaluable improvement for users, if it were to be integrated with the transcriptomics studies. However, to date there are still few datasets in which data from the methylome and transcriptome for the same specimens are available. The Additional Analyses section can be also expanded to include additional tools that can aid the researcher to gather valuable information about a gene(s) of interest. Potential tools of interest include Touchstone (https://clue.io/touchstone) and DEPMAP (https://depmap.org/portal/; ref. 19).

Our capacity to accrue empirical data is no longer the bottleneck in cancer research. With the appropriate bioinformatics expertise and access to public OMICs datasets, a hypothesis can be tested, refined, and the molecular mechanism of candidate genes can be predicted *in silico*. This strategy raises considerably the success rate and biomedical impact of cancer research projects. CANCERTOOL brings unique visualization and representation capabilities to cancer research teams that lack the personnel or the expertise to perform bioinformatics analyses, thus complementing and enriching the repertoire of available web tools.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: A.R. Cortazar, V. Torrano, A.M. Aransay, A. Carracedo
Development of methodology: A.R. Cortazar
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): A.R. Cortazar, V. Torrano, E. Guruceaga, S. Vicent
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A.R. Cortazar, V. Torrano, N. Martín-Martín, R.R. Gomis, I. Apaolaza, F.J. Planes, A.M. Aransay, A. Carracedo
Writing, review, and/or revision of the manuscript: A.R. Cortazar, A.M. Aransay, A. Carracedo
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): A.R. Cortazar
Study supervision: A.M. Aransay, A. Carracedo
Other (tool testing and review): V. Torrano, N. Martín-Martín, A. Caro-Maldonado, L. Camacho, I. Hermanova, L.F. Lorenzo-Martín, R. Caloto, I. Apaolaza, V. Quesada, S. Vicent, X.R. Bustelo, F.J. Planes, A.M. Aransay, A. Carracedo
Other (cosupervised L. Camacho's work): A. Gomez-Muñoz
Other (cosupervised I. Hermanova's work): J. Trka

Acknowledgments

Apologies to those whose related publications were not cited due to space limitations. We are grateful to Iñaki Lazaro for the design of the tumor type logos, Evarist Planet and Antoni Berenguer for insightful discussions, and the Carracedo lab for valuable input. V. Torrano is funded by Fundación Vasca de Innovación e Investigación Sanitarias, BIOEF (BIO15/CA/052), the AECC J.P. Bizkaia and the Basque Department of Health (2016111109). The work of A. Carracedo is supported by the Basque Department of Industry, Tourism and Trade (Etortek) and the Department of Education (IKERTALDE IT1106-16, also participated by A. Gomez-Muñoz), the BBVA Foundation, the MINECO [SAF2016-79381-R (FEDER/EU)]; Severo Ochoa Excellence Accreditation SEV-2016-0644; Excellence Networks (SAF2016-81975-REDT), European Training Networks Project (H2020-MSCA-ITN-308 2016 721532), the AECC IDEAS16 (IDEAS175CARR), and the European Research Council (Starting Grant 336343, PoC 754627). CIBERONC was cofunded with FEDER funds. The work of A. Aransay is supported by the Basque Department of Industry, Tourism and Trade (Etortek and Elkartek Programs), the Innovation Technology Department of Bizkaia County, CIBERehd Network, and Spanish MINECO the Severo Ochoa Excellence Accreditation (SEV-2016-0644). I. Apaolaza is funded by a Basque Government predoctoral grant (PRE_2017_2_0028). X.R. Bustelo is supported by grants from the Castilla-León Government (BIO/SA01/15, CSI049U16), Spanish Ministry of Economy and Competitiveness (MINECO; SAF2015-64556-R), Worldwide Cancer Research (14-1248), Ramón Areces Foundation, and the Spanish Society against Cancer (GC16173472GARC). Funding from MINECO to X.R. Bustelo is partially contributed by the European Regional Development Fund. The work of F.J. Planes is supported by the MINECO (BIO2016-77998-R) and ELKARTEK Programme of the Basque Government (KK-2016/00026).

Received June 1, 2018; revised August 2, 2018; accepted September 6, 2018; published first September 19, 2018.

Published OnlineFirst September 19, 2018; DOI: 10.1158/0008-5472.CAN-18-1669

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.
2. GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017; 390:1151–210.
3. Haber DA, Gray NS, Baselga J. The evolving war on cancer. Cell 2011; 145:19–24.
4. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20.
5. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
6. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004;6:1–6.
7. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401–4.
8. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6:pl1.
9. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185–93.
10. Vallejo A, Perurena N, Guruceaga E, Mazur PK, Martinez-Canarias S, Zandueta C, et al. An integrative approach unveils FOSL1 as an oncogene vulnerability in KRAS-driven lung and pancreatic cancer. Nat Commun 2017;8:14294.
11. Dunn OJ. Estimation of the medians for dependent variables. Ann Math Statist 1959;30:192–7.
12. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian J Stat 1979;6:65–70.
13. Edgington ES. An additive method for combining probability values from independent experiments. J Psychol 1972;80:351–63.
14. R Core Development Team. R: A language and environment for statistical computing. Vienna, Austria: R Core Development Team; 2011.
15. Wickham H. ggplot2: Elegant graphics for data analysis. New York, NY: Springer Publishing Company; 2009.
16. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 1966;50:163–70.
17. Hintze JL, Nelson RD. Violin plots: a box plot-density trace synergism. American Statistician 1998;52:181–4.
18. Valcarcel L, Macchia A, Martin N, Cortazar AR, Schaub A, Pujana M, et al. Integrative analysis of transcriptomics and clinical data uncovers the tumor-suppressive activity of MITF in prostate cancer. Cell Death Dis 2018;9:1041. doi: 10.1038/s41419-018-1096-6.
19. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. Cell 2017;170:564–76.
20. Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res 2006;66:10292–301.
21. Chitale D, Gong Y, Taylor BS, Broderick S, Brennan C, Somwar R, et al. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. Oncogene 2009;28:2773–83.
22. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest 2004;113:913–23.
23. Lu X, Wang ZC, Iglehart JD, Zhang X, Richardson AL. Predicting features of breast cancer with gene expression patterns. Breast Cancer Res Treat 2008;108:191–201.
24. Jorissen RN, Gibbs P, Christie M, Prakash S, Lipton L, Desai J, et al. Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage B and C colorectal cancer. Clin Cancer Res 2009;15:7642–51.
25. Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R, et al. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res 2012;72: 100–11.
26. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature 2012;487:239–43.
27. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486:346–52.
28. Kemper K, Versloot M, Cameron K, Colak S, de Sousa e Melo F, de Jong JH, et al. Mutations in the Ras-Raf Axis underlie the prognostic value of CD133 in colorectal cancer. Clin Cancer Res 2012;18:3132–41.
29. Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 2008; 14:822–7.
30. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A 2004;101:811–6.
31. Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 2005;7:R953–64.
32. Laibe S, Lagarde A, Ferrari A, Monges G, Birnbaum D, Olschwang S, et al. A seven-gene signature aggregates a subgroup of stage II colon cancers with stage III. OMICS 2012;16:560–5.
33. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell 2010;18:11–22.
34. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med 2013;10: e1001453.
35. Wilkerson MD, Yin X, Walter V, Zhao N, Cabanski CR, Hayward MC, et al. Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. PLoS One 2012;7:e36530.
36. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–9.
37. Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. Int J Cancer 2014;134:552–62.
38. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, et al. Integrative molecular concept modeling of prostate cancer progression. Nat Genet 2007;39:41–51.
39. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell 2005;8:393–406.

Cancer Research
The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

AACR American Association for Cancer Research

CANCERTOOL: A Visualization and Representation Interface to Exploit Cancer Datasets

Ana R. Cortazar, Veronica Torrano, Natalia Martín-Martín, et al.

Cancer Res 2018;78:6320-6328. Published OnlineFirst September 19, 2018.