PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx

Mohamed Mounir[1☯], Marta Lucchetta[1☯], Tiago C. Silva[2], Catharina Olsen[3,4], Gianluca Bontempi[3,4], Xi Chen[5,6], Houtan Noushmehr[2,7], Antonio Colaprico[3,4,6]*, Elena Papaleo[1,8]*

1 Computational Biology Laboratory, Danish Cancer Society Research Center, Copenhagen, Denmark, **2** Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil, **3** Interuniversity Institute of Bioinformatics in Brussels (IB)2, Brussels, Belgium, **4** Machine Learning Group (MLG), Department d'Informatique, Université libre de Bruxelles (ULB), Brussels, Belgium, **5** Sylvester Comprehensive Cancer Center, Miami, Florida, United States of America, **6** Division of Biostatistics, Department of Public Health Science, University of Miami Miller School of Medicine, Miami, Florida, United States of America, **7** Department of Neurosurgery, Henry Ford Hospital, Detroit, Michigan, United States of America, **8** Translational Disease Systems Biology, Faculty of Health and Medical Sciences, Novo Nordisk Foundation Center for Protein Research University of Copenhagen, Copenhagen, Denmark

☯ These authors contributed equally to this work.
* axc1833@med.miami.edu (AC); elenap@cancer.dk (EP)

OPEN ACCESS

Citation: Mounir M, Lucchetta M, Silva TC, Olsen C, Bontempi G, Chen X, et al. (2019) New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS Comput Biol 15(3): e1006701. https://doi.org/10.1371/journal.pcbi.1006701

Editor: Edwin Wang, University of Calgary Cumming School of Medicine, CANADA

Received: June 25, 2018

Accepted: December 10, 2018

Published: March 5, 2019

Data Availability Statement: Data are available through the github repository for examples used in the paper: https://github.com/ELELAB/TCGAbiolinks_examples and the code is available in: https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html; https://github.com/ELELAB/TCGAbiolinks; and https://github.com/BioinformaticsFMRP/TCGAbiolinks/.

Funding: The project was supported by a KBVU Pre-graduate scholarship 2017 to M.L. in EP's group, as well as by the LEO foundation grant

Abstract

The advent of Next-Generation Sequencing (NGS) technologies has opened new perspectives in deciphering the genetic mechanisms underlying complex diseases. Nowadays, the amount of genomic data is massive and substantial efforts and new tools are required to unveil the information hidden in the data. The Genomic Data Commons (GDC) Data Portal is a platform that contains different genomic studies including the ones from The Cancer Genome Atlas (TCGA) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiatives, accounting for more than 40 tumor types originating from nearly 30000 patients. Such platforms, although very attractive, must make sure the stored data are easily accessible and adequately harmonized. Moreover, they have the primary focus on the data storage in a unique place, and they do not provide a comprehensive toolkit for analyses and interpretation of the data. To fulfill this urgent need, comprehensive but easily accessible computational methods for integrative analyses of genomic data that do not renounce a robust statistical and theoretical framework are required. In this context, the *R/Bioconductor* package *TCGAbiolinks* was developed, offering a variety of bioinformatics functionalities. Here we introduce new features and enhancements of *TCGAbiolinks* in terms of i) more accurate and flexible pipelines for differential expression analyses, ii) different methods for tumor purity estimation and filtering, iii) integration of normal samples from other platforms iv) support for other genomics datasets, exemplified here by the TARGET data. Evidence has shown that accounting for tumor purity is essential in the study of tumorigenesis, as these factors promote confounding behavior regarding differential expression analysis. With this in mind, we implemented these filtering procedures in *TCGAbiolinks*. Moreover, a limitation of some of the TCGA datasets is the unavailability or paucity of

number LF17006. The calculations described in this paper were performed using the DeiC National Life Science Supercomputer at DTU. This work has also been supported by the BridgeIRIS project, funded by INNOVIRIS, Region de Bruxelles Capitale, Brussels, Belgium, and by GENGISCAN: GENomic profiling of Gastrointestinal Inflammatory-Sensitive CANcers, Belgian FNRS PDR (T100914F to AC, CO and GB), by institutional support from Henry Ford Health System (HN), and by the São Paulo Research Foundation (FAPESP) (2016/01389-7 to TCS & HN and 2015/07925-5 to HN). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

corresponding normal samples. We thus integrated into *TCGAbiolinks* the possibility to use normal samples from the Genotype-Tissue Expression (GTEx) project, which is another large-scale repository cataloging gene expression from healthy individuals. The new functionalities are available in the *TCGAbiolinks* version 2.8 and higher released in *Bioconductor* version 3.7.

Author summary

The advent of Next-Generation Sequencing (NGS) technologies has been generating a massive amount of data which require continuous efforts in developing and maintain computational tool for data analyses. The Genomic Data Commons (GDC) Data Portal is a platform that contains different cancer genomic studies. Such platforms have often the primary focus on the data storage and they do not provide a comprehensive toolkit for analyses. To fulfil this urgent need, comprehensive but accessible computational protocols that do not renounce a robust statistical framework are thus required. In this context, we here present the new functions of the *R/Bioconductor* package *TCGAbiolinks* to improve the discovery of differentially expressed genes in cancer and tumor (sub)types, include the estimate of tumor purity and tumor infiltrations, use normal samples from other platforms and support more broadly other genomics datasets.

This is a *PLOS Computational Biology* Software paper.

Introduction

Cancer is among the leading causes of mortality worldwide. It is a complex disease where multiple different mechanisms are at play all at once. This complexity also arises from the fact that cancer is extremely heterogeneous and can exist in distinct forms where each cancer type or subtype can be characterized by different molecular profiles with possible consequences on treatment and prognosis for the patient [1,2]. Advances in next-generation sequencing are currently making a massive amount of data available via the profiling of samples from cancer patients [3–7].

In this context, numerous large-scale studies have been conducted using state-of-the-art genome analysis technologies. One of the most important examples is The Cancer Genome Atlas (TCGA), which started in 2006 as a pilot project aiming to collect and conduct analyses on an unprecedented amount of clinical and molecular data including over 33 tumor types spanning over 11,000 patients. This project has subsequently generated more than 2.5 petabytes of publicly available data over the past decade [8,9]. Publicly funded by The National Institute of Health (NIH), TCGA has made numerous discoveries regarding genomic and epigenomic alterations that are candidate drivers for cancer development. This was achieved through the creation of an "atlas" and by applying large-scale genome-wide sequencing and multidimensional analyses. These efforts have significantly contributed to high-quality oncology studies, either led by the TCGA research network or other independent researchers [10],

which recently culminated in 27 original publications from the Pan-Cancer TCGA initiative [11]. In 2016, TCGA was moved under the umbrella of the broader repository Genomic Data Commons (GDC) Data Portal [12] together with other studies.

TCGA offers two versions of public data: legacy and harmonized. The legacy data is an unmodified collection of data that was previously maintained by the Data Coordinating Center (DCC) using GRCh36 (hg18) and GRCh37 (hg19) as genome reference assemblies. On the other hand, the harmonized version provides data that has been fully harmonized using GRCh38 (hg38) as a reference genome available through the GDC portal.

Many tools have been developed to interface with TCGA data [13–25] and to help with the aggregation, pre- and post-processing of the datasets. Among them, *TCGAbiolinks* was developed as an *R/Bioconductor* package to address the challenges of comprehensive analyses of TCGA data [19,20,26]. Software packages such as *TCGAbiolinks* regularly require enhancements and revisions in light of new biological or methodological evidence from the literature or new computational requirements imposed by the platforms where the data are stored.

For example, it is well-recognized that the tumor microenvironment also includes non-cancerous cells of which a large proportion are immune cells or cells that support blood vessels and other normal cells [27,28]. These components can ultimately alter the outcome of genomic analyses and the biological interpretation of the results. Recently, an extensive effort was made to systematically quantify tumor purity with a variety of diverse methods integrated into a consensus approach across TCGA cancer types [29], which the tools for analyses of TCGA data should employ.

Other cancer genomic initiatives have been following the TCGA model, such as Therapeutically Applicable Research to Generate Effective Treatments (TARGET), which is an NCI-funded project conducting a large-scale study that seeks to unravel novel therapeutic targets, biomarkers, and drug targets in childhood cancers by comprehensive molecular characterization and understanding of the genomic landscape in pediatric malignancies [30]. Comprehensive support for the analyses of different genomic datasets with the same workflow is thus essential for both reproducibility and harmonization of the results.

Lastly, it is common practice to use adjacent tissue showing normal characteristics at a macroscopic or histological level as a control. This advantageous practice concerning time-efficiency and reduction of patient-specific bias is based on the assumption that these samples are truly normal. Nevertheless, a tissue that is in the vicinity of or adjacent to a highly genetically abnormal tumor is likely to show cancer-related molecular aberrations [31], biasing the comparison. Moreover, circulating biomolecules, originating from cancer cells, can be taken in by the surrounding normal-like cells and alter their gene expression and processes. TCGA includes non-tumor samples from the same cancer participants. Furthermore, the pool of TCGA normal samples is often limited or lacking in TCGA projects. In this context, initiatives such as *Recount* [32], *Recount2* [33] and *RNASEQDB* [34] where TCGA data were integrated with normal healthy samples from the Genotype-Tissue Expression (GTEx) project [35] have the potential to boost the comparative analyses especially for those TCGA datasets where normal samples are underrepresented or unavailable.

In light of recent discoveries on the impact of tumor purity quantification on the samples under investigation [29], the need for a more substantial amount of normal samples [33], as well as the implementation of robust and statistically sound workflows for differential expression analyses [36,37] and exploration of potential sources of batch effects [38], we present new key features and enhancements that we implemented in *TCGAbiolinks* version 2.8 and higher.

Results

Overview of *TCGAbiolinks*

For the sake of clarity, we will briefly introduce the main functions of *TCGAbiolinks* that are extensively discussed in the original publication and a recently published workflow [19,20]. We advise referring directly to these publications and to the vignette on *Bioconductor* for more details about the basic functionalities.

The data retrieval is handled by the three main *TCGAbiolinks* functions: *GDCquery*, *GDCdownload* and *GDCprepare* and allows the user to interface with three main platforms: i) TCGA, ii) TARGET and, iii) The Cancer Genome Characterization Initiative (CGCI) (https://ocg.cancer.gov/programs/cgci). *TCGAbiolinks* also allows the user to interface with different -omics data including genomics and transcriptomics, clinical and pathological data, information on drug treatments, and subtypes.

GDCprepare allows the user to prepare the gene expression data for downstream analyses. This step is done by restructuring the data into a SummarizedExperiment (SE) object [39] that is easily manageable and integrable with other *R/Bioconductor* packages or just as a dataframe for other forms of data manipulation, which the user can operate even decoupled from the *TCGAbiolinks* package.

Moreover, *TCGAbiolinks* offers the option to apply normalization methods with the function *TCGAanalyze_Normalization* adopting the *EDASeq* protocol [40], to apply between-lane normalization to adjust for distributional differences between samples or within-lane normalization (to account for differences in GC content and gene length).

To guide result interpretation, the *TCGAvisualize* function allows the user to generate the plots required for a comprehensive view of the analyzed data using mostly the *ggplot2* package that has incremental layer options (such as principal component analysis, pathway enrichment analysis etc.) [41].

We extended *TCGAbiolinks* with new functionalities and methods that could boost the analyses of genomic data while at the same time not necessarily limiting these functionalities to just the TCGA initiative.

Towards a more generalized analyses of genomic data in GDC

TCGAbiolinks was initially conceived to interact with TCGA data, but the same workflow could be in principle extended to other datasets if the functions to handle their differences in formats and data availability are properly handled. Thus, we worked to support the SE format for other GDC datasets, such as the ones from the TARGET consortium which is included in *TCGAbiolinks version 2.8*. The SE object provides the advantage of collecting clinical information on the samples (such as patient gender, age and treatments) and on genes (ENSEMBL and ENTREZ IDs). One of the major problems in the study of genomic data is that they are often stored in unconnected silos which can lead to the of stalling of advancements in the analyses [42]. The design of the *GDCprepare* function of *TCGAbiolinks* thus nicely fulfills the need for standardized and harmonized ways to process data from different genomics initiatives which could find common storage in the GDC portal. Moreover, we provide the possibility to integrate data from external sources and carry out joint analyses with the GDC dataset (see the new *TCGAbatch_correction* function below).

Handling batch corrections in *TCGAbiolinks*: *TCGAbatch_Correction*

High-throughput sequencing and other -omics experiments are subject to unwanted sources of variability due to the presence of hidden variables and heterogeneity. Samples are processed

through different protocols, depending on the practices followed by each independent laboratory, involving time factors and multiple people orchestrating the genomic experiments. Known as batch effects, these sources of heterogeneity can have severe impacts on the results by statistically or biologically compromising the validity of the research [38,43,44].

Here, we created the *TCGAbatch_Correction* function to address and correct for different potential sources of batch effects linked to TCGA gene expression data using the *sva* package in R [38]. The *sva* package provides a framework for removing artifacts either by (i) estimating surrogate variables that introduce unwanted variability into high-throughput, high-dimensional datasets or (ii) using the *ComBat* function that employs an empirical Bayesian framework to remove batch effects related to known sources [44]. Modeling for known batch effects significantly helps to improve results by stabilizing error rates and reducing dependence on surrogates.

In this context, *TCGAbatch_Correction* takes GDC gene expression data as input, extracts all the needed metadata by parsing barcodes, corrects for a user-specified batch factor, and also adjusts for any selected cofactor. In cases where the investigator is not interested in correcting for batch effects with *ComBat* or this step is discouraged for the downstream analyses, the *voom* (an acronym for variance modeling at the observational level) transformation can be applied to carry out normal-based statistics on RNA-Seq gene counts [36] (see below).

The *TCGAbatch_Correction* function also generates plots to compare the parametric estimates for the distribution of batch effects across genes and their kernel estimates. Moreover, the so-called Q-Q plots can be produced showing the empirical data of ranked batch effects on each gene compared to their parametric estimate. Before applying batch effect corrections, one should investigate if there is any evidence of extreme differences between the kernel and the parametric estimates. Such differences can show up as bimodality or severe skewness and are due to the inability of the parametric estimation to pick up the empirical kernel behavior (an example is provided in the case study on breast cancer below and is discussed in Fig 1).

Additionally, to make TCGA data useful in a broader context, we included the possibility of integrating data from external sources or unpublished data in the context of publicly available datasets such as the ones in the GDC portal. To reach this goal, we have provided the possibility within the *TCGAbatch_Correction* function to integrate gene expression data from external sources (e.g GEO or unpublished datasets) and obtain a merged dataframe that can be used for further analysis within the *TCGAbiolinks* pipeline such as differential expression analysis. Nevertheless, we recommend the user to proceed with extreme caution with regards to the downstream analyses and to include the proper steps for batch corrections and harmonization of the data when they come from different sources. It is also important to rely on data that have been collected with the same technique and possibly the same instrument.

We provide an example for illustrative purposes only to handle the integration of datasets from external sources with TCGA data. The *TCGAbatch_Correction* function can be used to correct the integrated data for a common batch factor. In this example, we integrated the TCGA Lung Adenocarcinoma (LUAD) with the GEO dataset GSE60052 [45] where RNA-seq data are available for 79 samples from Small Cell Lung Cancer (SCLC) tissues and 7 normal controls. We restricted our analysis to only tumor samples in both datasets since there were no clear annotations for the normal samples on the GEO dataset. We queried, downloaded, and pre-processed the TCGA-LUAD data according to the workflow used in case study 1 and 2 (see below). We log2-transformed the TCGA data to make them comparable with the GEO data, which were released as log-transformed values. We decided to correct the data according to the year when the sample was taken since it is the only factor in common and a suitable candidate to correct for technical variability in this example. We retrieved the sample year from the downloaded TCGA clinical data using the *GDCquery_clinic* function. The GEO clinical

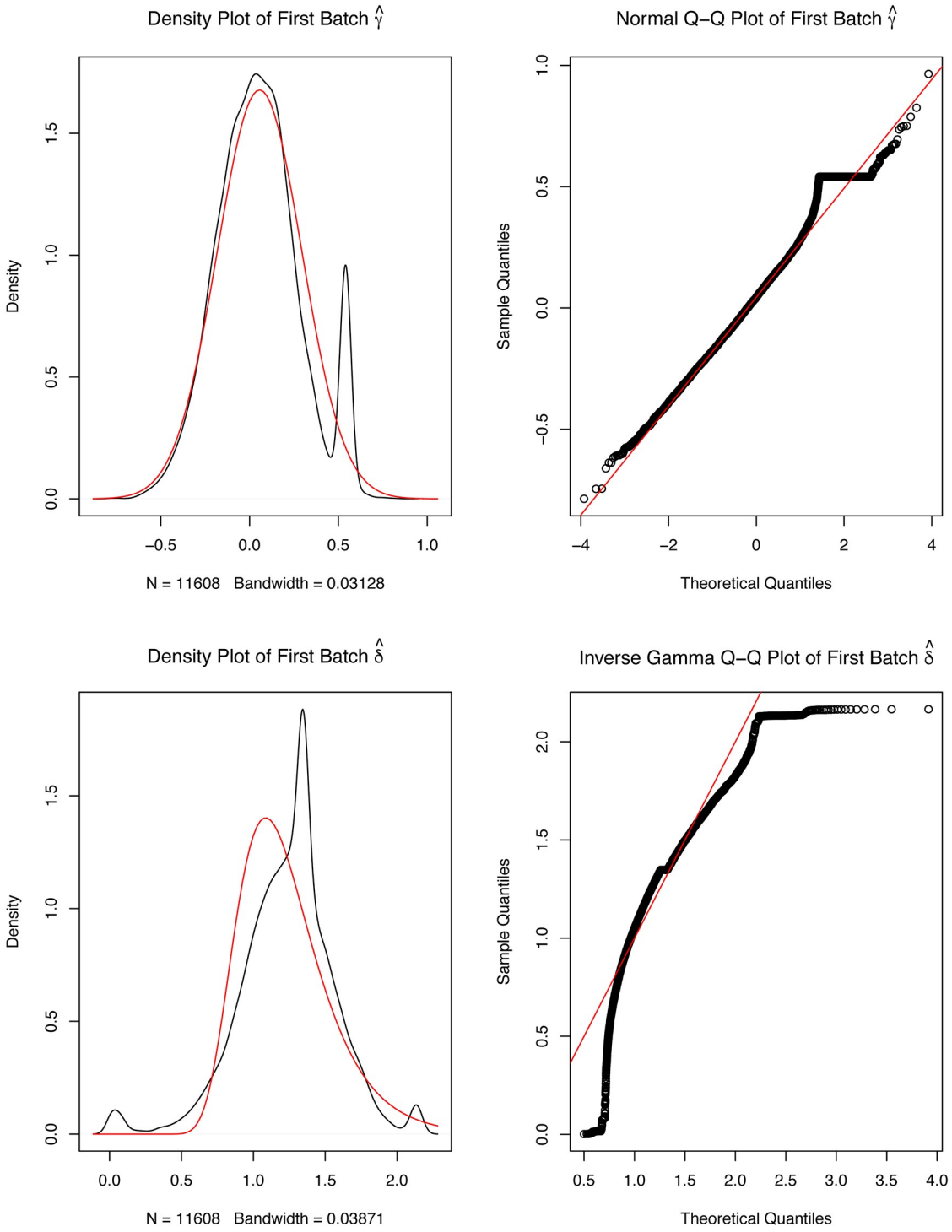

Fig 1. Example of the exploration of batch effects. Four plots generated by *ComBat* to correct for batch effects. For the left panel plots, the red lines are the parametric estimates, and the black lines are the kernel estimates for the distribution of effects across genes. The right panel shows Q-Q plots with the red line for the parametric estimate and the ordered batch effects for each gene (black points). The bottom plots show the analyses for the variances and the top plots refers to the means. Plots were generated for batches TSS E9 and E2 to avoid batches containing only one sample.

https://doi.org/10.1371/journal.pcbi.1006701.g001

data has been released as supplementary material to the original publication (Table S1 in [45]). In particular, we selected all the tumor samples taken from 2010 to 2012 (three batches in total) in both datasets. We also ensured that more than one sample was available for each batch. The tumor samples which fulfilled the chosen batch criterion were 50 and 21 in TCGA and GEO, respectively. Since TCGA includes 17400 and GEO 15711 genes, we selected only the features in common (15711) by converting the TCGA Ensembl IDs to gene names using the information stored in the SummarizedExperiment object, retrieved through the *rowData* function.

We then merged the two datasets and created the corresponding batch information. This information was then provided as input to the *TCGAbatch_Correction* function to produce the integrated year-corrected matrix. The script to reproduce this example is available in the GitHub repository associated to this publication (https://github.com/ELELAB/TCGAbiolinks_examples). We would like to stress the fact that this is just an example to show how the function works. In a real case study, the best course of action would be to process the external (GEO) data and the TCGA data through the same pipeline, starting from the external raw data and calculating the read count as it is done in the harmonized or legacy version of the TCGA data, depending on the dataset of interest for the comparison.

TCGA_MolecularSubtype

Although each cancer is believed to be a single disease, advances in the genomic field now indicate that each cancer type is much more heterogeneous than previously thought and that different subtypes can be identified. Bioinformatics applied to genomics data can enable a molecular understanding of the tumors across different cancer subtypes. Instead of binning all cases and patients into a single category, differentiating the intrinsic subtypes of each cancer has provided efficient, targeted, treatment strategies and prognoses. Cancer subtypes can be defined according to histology or molecular profiles. Tables with general annotations from the TCGA publications on classifications of the patients are provided by the *TCGAquery_subtype* function [19]. However, the format of these data is not so easy to navigate or integrate within other functions.

For this reason, we designed a new function *TCGA_MolecularSubtype* to retrieve information on manually curated molecular subtypes for a total of 24 cancer types (Table 1). Collectively, we have molecular subtype annotations for 7734 individuals. The function also allows fetching of the subtype information not only for each cancer type, but also for each TCGA barcode (i.e. for each individual sample). The information used to classify cancer subtypes is the one used (and most recently published) by the Pan-Cancer works from the TCGA consortium (http://bioinformaticsfmrp.github.io/TCGAbiolinks/subtypes.html#pancanceratlas_subtypes:_curated_molecular_subtypes). As an alternative, there is also the *PanCancerAtlas_subtypes* function. These new functions have the advantage that the data are manually curated from each TCGA cancer type marker paper and are thus up to date when a new paper from the TCGA research network is published and reported in https://gdc.cancer.gov/about-data/publications.

Recently, we showed the advantage of using these functions to have a curated matrix in one single place for all of the subtypes. In particular, it has been applied to identify associations between molecular subtypes and the stemness index [46] and the immune subtypes [47] of TCGA samples.

TCGAtumor_purity

The tumor microenvironment encompasses cellular and non-cellular units that play a critical role in the initiation, progression, and metastasis of the tumor [27,29,48–50].

Table 1. Information on molecular subtypes for TCGA cancer studies as provided by the *TCGA_MolecularSubtype* function.

TCGA Abbreviation	Cancer type	Number of samples	Subtypes Selected
ACC	Adrenocortical carcinoma	91	ACC.CIMP-high, ACC.CIMP-intermediate, ACC.CIMP-low
AML	Acute Myeloid Leukemia	187	AML.1, AML.2, AML.3, AML.4, AML.5, AML.6, AML.7
BLCA	Bladder Urothelial Carcinoma	129	BLCA.1, BLCA.2, BLCA.3, BLCA.4
BRCA	Breast invasive carcinoma	1218	BRCA.Basal, BRCA.Her2, BRCA.LumA, BRCA.LumB, BRCA.Normal
COAD	Colon adenocarcinoma	341	GI.CIN, GI.GS, GI.HM-indel, GI.HM-SNV
ESCA	Esophageal carcinoma	169	GI.CIN, GI.ESCC, GI.GS, GI.HM-indel, GI.HM-SNV
GBM	Glioblastoma multiforme	606	GBM_LGG.Classic-like, GBM_LGG.Codel, GBM_LGG.G-CIMP-high, GBM_LGG.G-CIMP-low, GBM_LGG.LGm6-GBM, GBM_LGG.Mesenchymal-like
HNSC	Head and Neck squamous cell carcinoma	279	HNSC.Atypical, HNSC.Basal, HNSC.Classical, HNSC.Mesenchymal
KICH	Kidney Chromophobe	66	KICH.Eosin.0, KICH.Eosin.1
KIRC	Kidney renal clear cell carcinoma	442	KIRC.1, KIRC.2, KIRC.3, KIRC.4
KIRP	Kidney renal papillary cell carcinoma	161	KIRP.C1, KIRP.C2a, KIRP.C2b, KIRP.C2c - CIMP
LGG	Brain Lower Grade Glioma	516	GBM_LGG.Classic-like, GBM_LGG.Codel, GBM_LGG.G-CIMP-high, GBM_LGG.G-CIMP-low, GBM_LGG.Mesenchymal-like, GBM_LGG.PA-like
LIHC	Liver hepatocellular carcinoma	196	LIHC.iCluster:1, LIHC.iCluster:2, LIHC.iCluster:3
LUAD	Lung adenocarcinoma	230	LUAD.1, LUAD.2, LUAD.3, LUAD.4, LUAD.5, LUAD.6
LUSC	Lung squamous cell carcinoma	178	LUSC.basal, LUSC.classical, LUSC.primitive, LUSC.secretory
OVCA	Ovarian serous cystadenocarcinoma	489	OVCA.Differentiated, OVCA.Immunoreactive, OVCA.Mesenchymal, OVCA.Proliferative
PCPG	Pheochromocytoma and Paraganglioma	178	PCPG.Cortical admixture, PCPG.Pseudohypoxia, PCPG.Wnt-altered
PRAD	Prostate adenocarcinoma	333	PRAD.1-ERG, PRAD.2-ETV1, PRAD.3-ETV4, PRAD.4-FLI1, PRAD.5-SPOP, PRAD.6-FOXA1, PRAD.7-IDH1, PRAD.8-other
READ	Rectum adenocarcinoma	118	GI.CIN, GI.GS, GI.HM-indel, GI.HM-SNV
SKCM	Skin Cutaneous Melanoma	333	SKCM.-, SKCM.BRAF_Hotspot_Mutants, SKCM.NF1_Any_Mutants, SKCM.RAS_Hotspot_Mutants, SKCM.Triple_WT
STAD	Stomach adenocarcinoma	383	GI.CIN, GI.EBV, GI.GS, GI.HM-indel, GI.HM-SNV
THCA	Thyroid carcinoma	496	THCA.1, THCA.2, THCA.3, THCA.4, THCA.5
UCEC	Uterine Corpus Endometrial Carcinoma	538	UCEC.CN_HIGH, UCEC.CN_LOW, UCEC.MSI, UCEC.POLE
UCS	Uterine Carcinosarcoma	57	UCS.1, UCS.2

https://doi.org/10.1371/journal.pcbi.1006701.t001

An important concept to remember from the TME definition is that tumor purity is described as the proportion of carcinoma cells in a tumor sample. In previous times, tumor purity used to be estimated through visual inspection with the assistance of a pathologist and by image analysis. Nowadays, with the advent of computational methods and the use of genomic features such as somatic mutations, DNA methylation, and somatic copy-number variation (CNV), it is feasible to estimate tumor purity [27].

To account for tumor purity in the *TCGAbiolinks* workflow, we designed the *TCGAtumor_purity* function that filters data according to one of the following five methods: i) ESTIMATE (Estimation of Stromal and Immune cells in Malignant Tumor tissues using Expression data) [49]; ii) ABSOLUTE to infer tumor purity from the analysis of somatic DNA aberrations [50]; iii) LUMP (Leukocytes Unmethylation) that uses the average of 44 detected non-methylated immune-specific CpG site; iv) IHC, that uses hematoxylin- and eosin–stained slides, provided by the Nationwide Children's Hospital Biospecimen Core Resource, which

are processed using image analysis techniques to generate a tumor purity estimate; v) Consensus measurement of Purity Estimation (CPE), a consensus estimate from the four methods mentioned above [29]. CPE is calculated as the median purity level after normalization of the values from the four methods and correcting for the means and standard deviations and it is the default option of the *TCGAtumor_purity* function.

TCGAanalyze_DEA extension

We revised and expanded the pre-existing *TCGAbiolinks* function *TCGAanalyze_DEA* that performs differential expression analysis (DEA) by calling the commonly used R package, *edgeR* [37]. In the former version of *TCGAbiolinks*, only a pairwise approach (for example, control versus case) was applied to a matrix of count data and samples to extract differentially expressed genes (DEGs). More specifically, the former *TCGAanalyze_DEA* function implemented two options: (i) the *exactTest* framework for a simple pairwise comparison or (ii) the *GLM* (Generalized Linear Model) where a user faces a more complex experimental design involving multiple factors. However, in the latter case, the design of the function allowed the user to provide arguments only for case and control thereby being incompatible with multifactor experiments, for which GLM methods are particularly suited [51]. We thus implemented a different design to improve the functionality of *TCGAanalyze_DEA* by providing the ability to analyze RNA-Seq data in a more general and comprehensive way. The user is now able to apply *edgeR* with a more sophisticated design matrix and to use the *limma-voom* method, an emerging gold standard for RNA-Seq data [52]. Furthermore, modeling multifactor experiments and correcting for batch effects related to TCGA samples is now an option in the updated version of *TCGAanalyze_DEA*. The new arguments for the function allow to use different sources of batch effects in the design matrix, such as the plates, the TSS (Tissue Source Site), the year in which the sample was taken and the patient factor in the cases of paired normal and tumor samples. Moreover, an option is provided to apply two different pipelines to the study of paired or unpaired samples, namely *limma-voom* and *limma-trend* pipelines. A contrast formula is provided to determine coefficients and design contrasts in a customized way, as well as the possibility to model a multifactor experimental design. In particular, the model formula for the *edgeR* pipeline is designed so that the intercept is set to 0 when there are multiple conditions (such as the molecular subtypes) or contrasts to be explored, following the recommendation of *edgeR* developers.

The function returns two types of objects: i) a table with DEGs containing logFC, logCPM, p-value, and FDR corrected p-values in cases of pairwise comparison for each gene, and/or ii) a list object containing multiple tables for DEGs according to each contrast specified in the *contrast.formula* argument.

TCGAquery_recount2

The *Recount* project was created as an online resource that comprises gene count matrices built from 8 billion reads using 475 samples gathered from 18 published studies [32]. This atlas of RNA-Seq count matrices improves the process of data acquisition and allows cross-study comparisons since all of the count matrices were produced from one single pipeline reducing batch effects and promoting alternative normalization. *Recount* was then extended to *Recount2* consisting of more than 4.4 trillion reads using 70,603 human RNA-seq samples from the Sequence Read Archive (SRA), GTEx, and TCGA that were uniformly processed, quantified with Rail-RNA [51], and included in the recent R*ecount2* interface [33].

For this reason, *TCGAquery_recount2* queries GTEx and TCGA data for all tissues available in the *Recount2* platform, providing the user with the flexibility to decide which tissue source to use for the calculations.

TCGAquery_recount2 integrates normal samples from GTEx and normal samples from TCGA. If the user wants to use GTEx alone as a source of normal samples, an *ad hoc* curation of the dataset will be needed before applying the functions for pre-processing of the data and downstream analyses with *TCGAbiolinks*.

Below, we illustrate two case studies as an example of the usage of the new functions and the interpretation of their results.

Case study 1—A protocol for pre-processing and differential expression analysis of TCGA-BRCA luminal subtypes

The TCGA Breast Invasive Carcinoma (BRCA) dataset is the ideal case study to illustrate the new functionalities of *TCGAbiolinks* (see Fig 2 for a workflow illustrating this case study and the new functions).

We carried out the query, download and pre-processing of the TCGA-BRCA RNA-Seq data through the GDC portal with a variation of the workflow suggested for the previous versions of the *TCGAbiolinks* software (see the script reported in https://github.com/ELELAB/TCGAbiolinks_examples). As an example, out of a possible 1222 BRCA samples available in the GDC portal, we restricted our analysis to 100 tumor (TP) samples and 100 normal (NT) samples respectively.

We constructed the SE object as the starting structure displaying information for both genes and samples with gene expression tables of HTSeq-based counts from reads harmonized and aligned to hg38 genome assembly. Afterwards, we applied an Array Array Intensity correlation (AAIC) to pinpoint samples with low correlation (0.6 threshold for this study) using *TCGAanalyze_Preprocessing*, which generates a count matrix ready to be used as input for the downstream analysis pipeline. In addition, we normalized the gene counts for GC-content using *TCGAanalyze_Normalization* adopting *EDASeq* protocol incorporated with *TCGAbiolinks*.

An exploratory data analysis (EDA) step is now possible within *TCGAbiolinks* to help to understand the quality of the data and to identify possible anomalies or cofounder effects. This can be done by estimating the presence of batch effects through the plots provided by the *ComBat* function, as described above. We can call the *TCGAbatch_Correction* function on a log2 transformed instance of the count matrix. For the sake of clarity, we used batch correction on TSS as a cofounder factor along with accounting for one covariate (cancer versus normal) and only two batches were retained. The results are reported in Fig 1.

According to the standard defined by the TCGA consortium, 60% tumor purity is the recommended threshold for analyses [29]. Thus, we applied a filtering step using the *TCGAtumor_purity* function of *TCGAbiolinks* whereby tumor samples that show a purity of less than 60% median CPE are discarded from the analysis. As a result, a total of 26 samples were discarded with the goal of reducing the confounding effect of tumor purity on genomic analyses.

We then applied the new *TCGAanalyze_DEA* function to exploit the power of generalized linear models beyond the control versus case scheme. As an illustrative case, we queried the PAM50 classification [52] for each of the samples through *TCGA_MolecularSubtype*. We identified 86 samples with information on subtypes. The output is then provided to the DEA method so the customizable *contrast.formula* argument can contain the formula for designing the contrasts. Beforehand, the data is normalized for GC-content, as explained above. As a final step, quantile filtering is applied with a cutoff of 25%, as suggested by the original

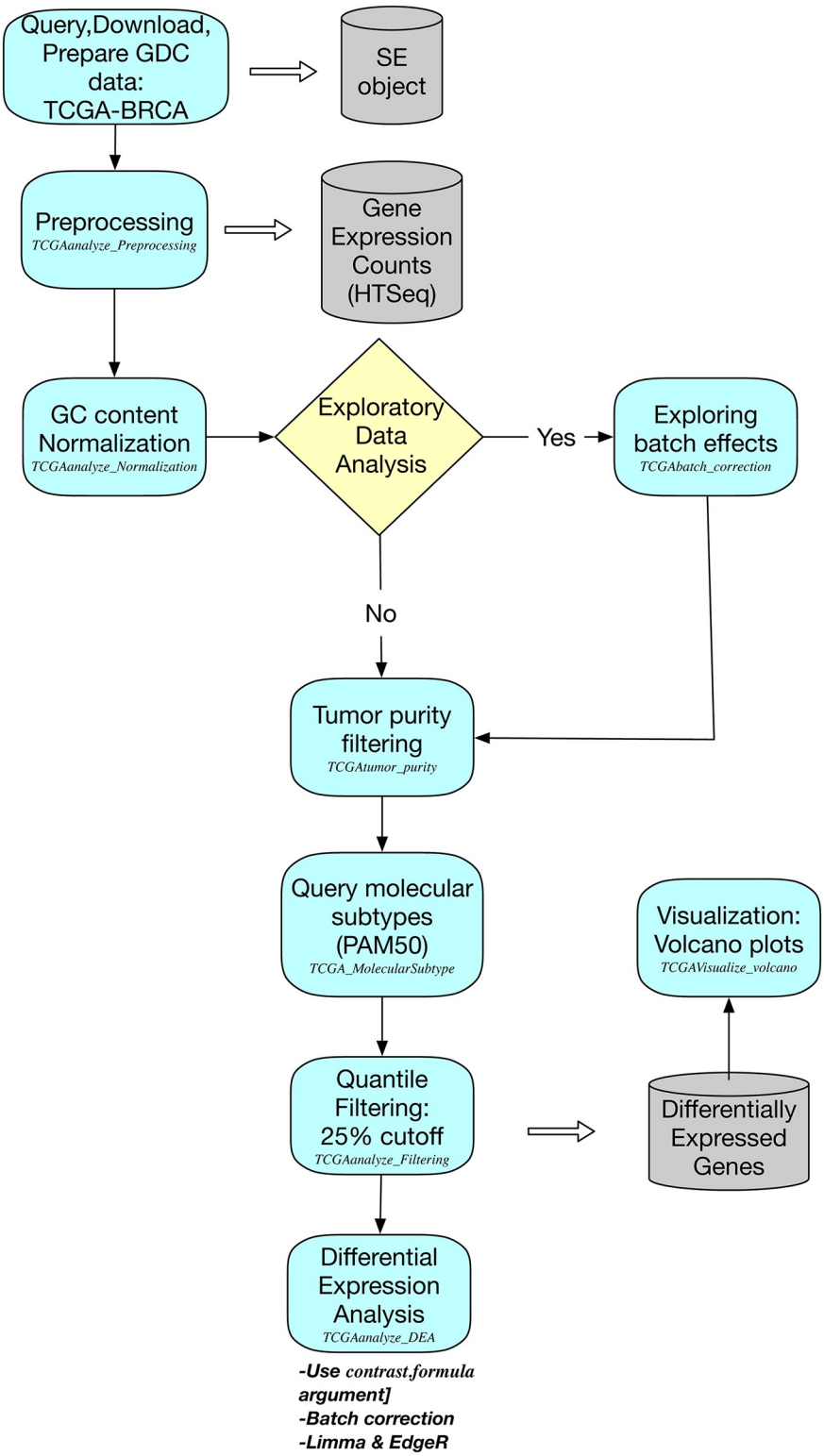

Fig 2. The workflow illustrates the steps and *TCGAbiolinks* functions to be used for case study 1 on TCGA-BRCA luminal subtypes.

https://doi.org/10.1371/journal.pcbi.1006701.g002

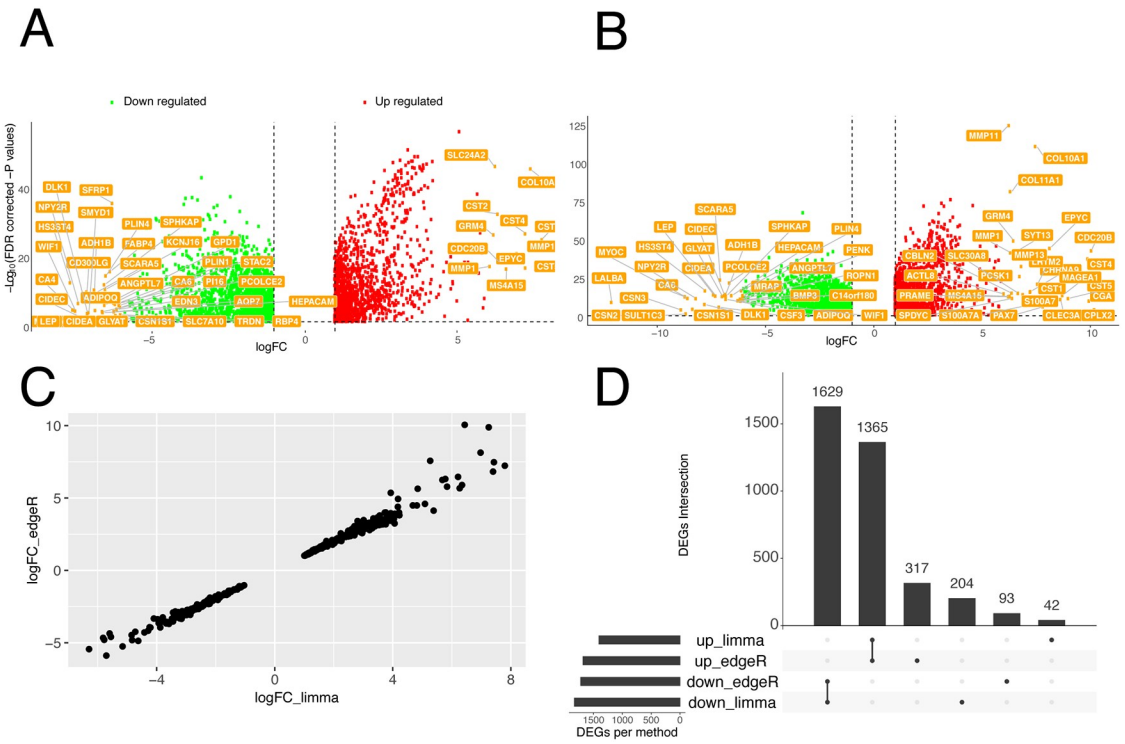

Fig 3. DEA analyses of TCGA-BRCA data comparing luminal subtypes with normal samples. A-B) Volcano plots are shown where only those genes with logFC higher than 6 or lower than -6 are labelled and only the significant up- or down-regulated genes are shown as dots. We carried out DEA using the *limma* (A) or *edgeR* pipelines (B) of *TCGAbiolinks*. C) The correlation plot between the logFC estimated by the two pipelines for the top 500 DE genes is shown. The genes discussed in the main text are highlighted in bold. D) The intersect between all the DE genes estimated by the two pipelines is shown using *UpSet*.

https://doi.org/10.1371/journal.pcbi.1006701.g003

TCGAbiolinks workflow. Within the *TCGAanalyze_DEA* function, it is also possible to perform a *voom* transformation of the count data, as detailed above. In Fig 3A, we show the results of the new implementation of the *TCGAanalyze_DEA* function as a volcano plot. The genes with highest logFC are shown (using logFC higher or lower than 6 in absolute value as a cut-off). We then compared these results to the ones produced using DEA as implemented in *edgeR* within the *TCGAanalyze_DEA* (see volcano plot in Fig 3B). We calculated the correlation between the top 500 DE genes identified by the two methods (Fig 3C) which resulted in a Pearson Correlation Coefficient higher than 0.9. We then quantitatively compared the results of the two methods calculating the intersect with *UpSetR* [53] (Fig 3D). The two methods are in good agreement showing 1629 and 1365 down- and up-regulated genes in common, which account for approximately 90% of the total DE genes. With both methods we identified up-regulated matrix metalloproteinases (such as MMP11 and MMP13) which are a class of enzyme known to be involved in cancer invasion and metastasis and have been linked to breast cancer outcomes [54]. We also identified different collagen proteins (such as COL10A1 and COL11A1) that are up-regulated in luminal versus normal breast cancer samples. Those proteins are important for the composition of the extracellular matrix (ECM). Changes in the amount or composition of the ECM have been considered a hallmark of tumor development [55]. COL11A1 and COL10A1 have recently been proposed as markers to discriminate between breast cancer and healthy tissues and could be helpful in the diagnosis of suspicious breast nodules [56].

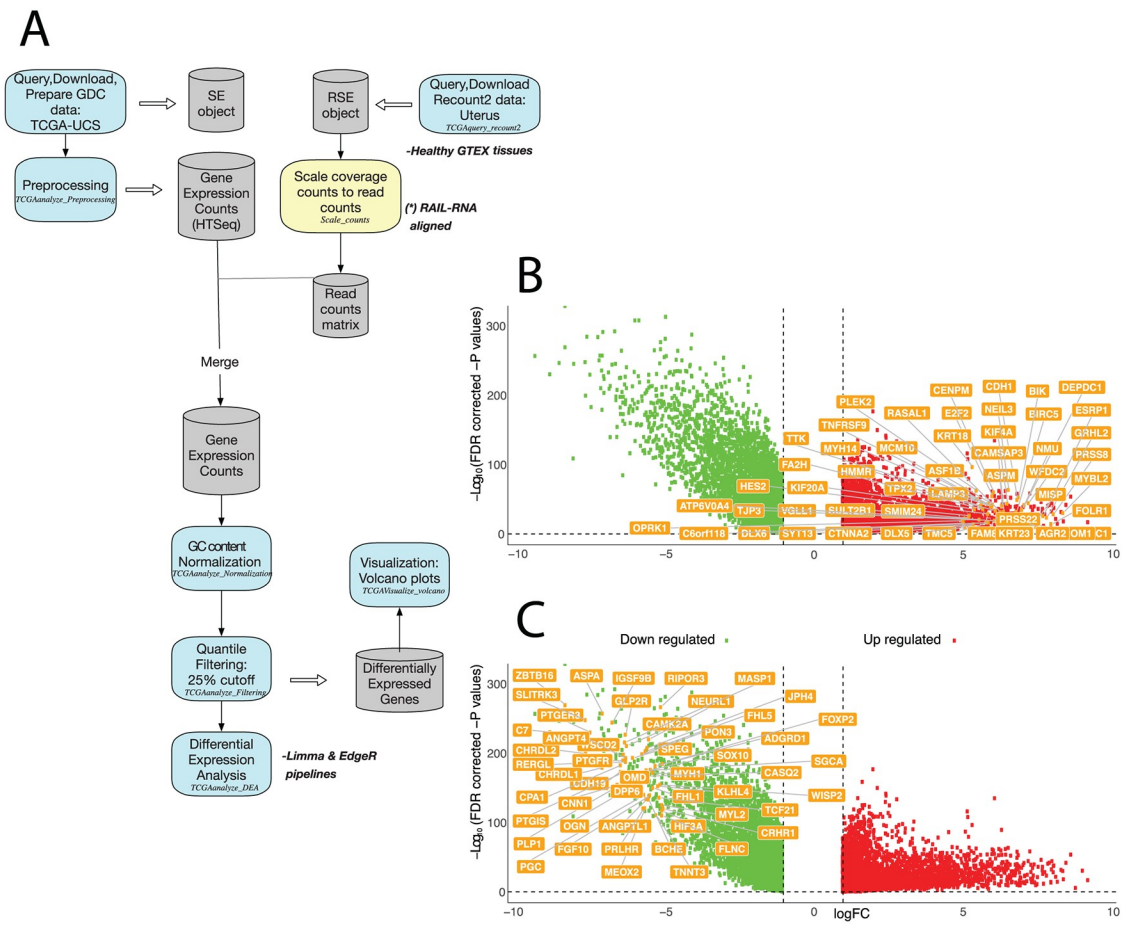

Fig 4. DE genes in uterine cancer compared to healthy uterine tissue samples. A) The workflow illustrates the steps and *TCGAbiolinks* functions to be used for this case study. B-C) In the volcano plot, the up-regulated genes with logFC higher than 5 (B) or the down-regulated genes with logFC lower than -5 (C) are shown as a result of DEA carried out using the *limma* pipeline comparing primary tumor samples from TCGA-UCS and normal uterine tissue samples from GTEx.

https://doi.org/10.1371/journal.pcbi.1006701.g004

Case study 2—Uterine cancer dataset exploiting Recount2

One issue that can be encountered when planning DEA of TCGA data is the fact that some projects on the GDC portal do not contain normal control samples for the comparison with the tumor samples. As explained previously, it is now possible to query data from the *Recount2* platform to increase the pool of normal samples and apply the DEA pipelines of *TCGAbiolinks* (see Fig 4A for a workflow).

For this case study, we used the TCGA Uterine Carcinosarcoma (UCS) dataset to illustrate this application. We queried, downloaded, and pre-processed the data using a similar workflow to our previous case study, and then GTEx healthy uterine tissues were used as a source of normal samples for DEA. Concerning the type of count data queried, it was similarly harmonized HTSeq counts and aligned to the hg38 genome assembly (see the script reported in https://github.com/ELELAB/TCGAbiolinks_examples). We used the *TCGAquery_recount2* function to download tumor and normal uterine samples from the *Recount2* platform as Ranged Summarized Experiment (RSE) objects.

Before engaging in DEA, one should keep in mind that the *Recount2* resource contains reads, some of them soft-clipped, aligned to *Gencode* version 25 hg38 using the splice-aware

Rail-RNA aligner. Moreover, the RSE shows coverage counts instead of standard read count matrices. Since most methods are adapted to read count matrices, there are some highly recommended transformations to perform before commencing with DEA. The user should extract sample metadata from RSE objects regarding read length and mapped read counts to pre-process the data. If one provides a target library size (40 million reads by default), coverage counts can be scaled to read counts usable for classic DEA methods according to Eq (1) (possibly with the need to round the counts since the result might not be of an integer type).

$$\sum_{i}^{n} \frac{coverage}{Read\ \ Length} * \frac{target}{mapped} \quad = \quad scaled\ \ read\ \ counts \tag{1}$$

The denominator is the sum of the coverage for all base-pairs of the genome which can be replaced by the Area under Curve (AUC) [57]. It is possible to use the function *scale_counts* from the *recount* package. After that, we merged the two prepared gene count matrices, normalized for GC-content and applied the quantile filtering with a 25% cut-off. The data were then loaded into the *TCGAanalyze_DEA* function for comparison of normal samples versus cancer samples using the *limma-voom* pipeline. Two volcano plots depicting the top up- and down-regulated genes are shown in Fig 4B and 4C, respectively. As an example, we identified the up-regulated gene ADAM28 in the UCS tumor samples when compared to the normal ones (logFC = 3.13, thus not shown in Fig 4B). ADAM28 belongs to the ADAM family of disintegrins and metalloproteinases which are involved in important biological events such as cell adhesion, fusion, migration and membrane protein shedding and proteolysis. They are often overexpressed in tumors and contribute to the promotion of cell growth and invasion [58]. Among the top up-regulated genes in UCS, we also identified other key players in cell adhesion such as the cadherin CDH1 [58] shown in Fig 4B.

Availability and future directions

The functions illustrated in this manuscript are now available in version 2.8 of *TCGAbiolinks* on *Bioconductor* version 3.7 (https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html *)*, as well as through the two Github repositories (https://github.com/ELELAB/TCGAbiolinks and https://github.com/BioinformaticsFMRP/TCGAbiolinks/).

In addition, we provide daily scientific advice to the Github community within the 'issues' forum (https://github.com/BioinformaticsFMRP/TCGAbiolinks/issues) to solve both software bugs and to provide new functionalities needed or requested by the Github community. This forum is also a place where *TCGAbiolinks* users can share and discuss their experience with their analyses with our team and/or other Github users.

The newly developed functions will for the first time allow users to fully appreciate the effect of using genuinely healthy samples or normal tumor-adjacent samples as a control as well as the benefits of correcting for the tumor purity of the samples. We provide a more robust and comprehensive workflow to carry out differential expression analysis with two different methods and a customizable design matrix, as well as the capability to handle batch corrections. Overall, this will provide the community with the possibility to use the same framework for vital analyses such as the benchmarking of differential expression methods.

(https://bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/extension.html).

Acknowledgments

The authors would like to thank Matteo Tiberti for fruitful discussions and comments and Lisa Cantwell for the professional scientific proofreading of the manuscript.

Author Contributions

Conceptualization: Elena Papaleo.

Data curation: Mohamed Mounir, Marta Lucchetta, Tiago C. Silva, Antonio Colaprico, Elena Papaleo.

Formal analysis: Mohamed Mounir, Marta Lucchetta, Elena Papaleo.

Funding acquisition: Gianluca Bontempi, Xi Chen, Houtan Noushmehr, Elena Papaleo.

Investigation: Marta Lucchetta, Elena Papaleo.

Methodology: Mohamed Mounir, Marta Lucchetta, Antonio Colaprico, Elena Papaleo.

Project administration: Elena Papaleo.

Resources: Elena Papaleo.

Software: Mohamed Mounir, Tiago C. Silva, Antonio Colaprico.

Supervision: Antonio Colaprico, Elena Papaleo.

Validation: Marta Lucchetta, Elena Papaleo.

Visualization: Mohamed Mounir, Marta Lucchetta, Elena Papaleo.

Writing – original draft: Elena Papaleo.

Writing – review & editing: Mohamed Mounir, Marta Lucchetta, Tiago C. Silva, Catharina Olsen, Gianluca Bontempi, Xi Chen, Houtan Noushmehr, Antonio Colaprico, Elena Papaleo.

References

1. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: A looking glass for cancer? Nat Rev Cancer. Nature Publishing Group; 2012; 12: 323–334. https://doi.org/10.1038/nrc3261 PMID: 22513401
2. Burrell R a, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013; 501: 338–45. https://doi.org/10.1038/nature12625 PMID: 24048066
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10: 57–63. https://doi.org/10.1038/nrg2484 PMID: 19015660
4. Nakagawa H, Wardell CP, Furuta M, Taniguchi H, Fujimoto A. Cancer whole-genome sequencing: present and future. Oncogene. Nature Publishing Group; 2015; 1–8. https://doi.org/10.1038/onc.2015.90 PMID: 25823020
5. Van Verk MC, Hickman R, Pieterse CMJ, Van Wees SCM. RNA-Seq: Revelation of the messengers. Trends Plant Sci. 2013; 18: 175–179. https://doi.org/10.1016/j.tplants.2013.02.001 PMID: 23481128
6. McGettigan PA. Transcriptomics in the RNA-seq era. Curr Opin Chem Biol. 2013; 17: 4–11. https://doi.org/10.1016/j.cbpa.2012.12.008 PMID: 23290152
7. LeBlanc VG, Marra MA. Next-Generation Sequencing Approaches in Cancer: Where Have They Brought Us and Where Will They Take Us? Cancers (Basel). 2015; 7: 1925–1958. PMID: 26404381
8. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. Nature Publishing Group; 2013; 45: 1113–1120. https://doi.org/10.1038/ng.2764 PMID: 24071849
9. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Wspolczesna Onkol. 2015; 1A: A68–A77. https://doi.org/10.5114/wo.2014.47136 PMID: 25691825
10. Hinkson I V., Davidsen TM, Klemm JD, Kerlavage AR, Kibbe WA. A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. Front Cell Dev Biol. 2017;5.

11. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell. Elsevier Inc.; 2018; 173: 283–285. https://doi.org/10.1016/j.cell.2018.03.042 PMID: 29625045

12. Grossman RL, Heath A, Murphy M. A Case for Data Commons: Toward Data Science as a Service. Comput Sci Eng. 2016; 18: 10–20. http://dx.doi.org/10.1109/MCSE.2016.92 PMID: 29033693

13. Samur MK. RTCGAToolbox: A New Tool for Exporting TCGA firehose data. PLoS One. 2014; 9. https://doi.org/10.1371/journal.pone.0106397 PMID: 25181531

14. Rodrigues João F Matias, Schmidt Thomas SB J T and von CM. TCGA-Assembler 2: Software Pipeline for Retrieval and Processing of TCGA/CPTAC Data. Bioinformatics. 2017; 0–0.

15. Chandran UR, Medvedeva OP, Barmada MM, Blood PD, Chakka A, Luthra S, et al. TCGA Expedition: A Data Acquisition and Management System for TCGA Data. PLoS One. 2016; 11: e0165395. https://doi.org/10.1371/journal.pone.0165395 PMID: 27788220

16. Cline MS, Craft B, Swatloski T, Goldman M, Ma S, Haussler D, et al. Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. Sci Rep. 2013; 3. https://doi.org/10.1038/srep02652 PMID: 24084870

17. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2015; 44: gkv1507-. https://doi.org/10.1093/nar/gkv1507 PMID: 26704973

18. Silva TC, Colaprico A, Olsen C, D'Angelo F, Bontempi G, Ceccarelli M, et al. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Research. 2016; 5: 1542. https://doi.org/10.12688/f1000research.8923.2 PMID: 28232861

19. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017; 45: W98–W102. https://doi.org/10.1093/nar/gkx247 PMID: 28407145

20. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. PeerJ Comput Sci. 2016; 2: e67. https://doi.org/10.7717/peerj-cs.67

21. Krasnov GS, Dmitriev AA, Melnikova N V., Zaretsky AR, Nasedkina T V., Zasedatelev AS, et al. CrossHub: A tool for multi-way analysis of the Cancer Genome Atlas (TCGA) in the context of gene expression regulation mechanisms. Nucleic Acids Res. 2016; 44: 1–11.

22. Deng M, Brägelmann J, Schultze JL, Perner S. Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. BMC Bioinformatics. BMC Bioinformatics; 2016; 17: 72. https://doi.org/10.1186/s12859-016-0917-9 PMID: 26852330

23. Wan Y-W, Allen GI, Liu Z. TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R. Bioinformatics. 2015; btv677-. https://doi.org/10.1093/bioinformatics/btv677 PMID: 26568634

24. Ryan M, Wong WC, Brown R, Akbani R, Su X, Broom B, et al. TCGASpliceSeq a compendium of alternative mRNA splicing in cancer. Nucleic Acids Res. 2016; 44: D1018–D1022. https://doi.org/10.1093/nar/gkv1288 PMID: 26602693

25. Zhang Z, Li H, Jiang S, Li R, Li W, Chen H, et al. A survey and evaluation of Web-based tools/databases for variant analysis of TCGA data. Brief Bioinform. 2018; 1–18.

26. Zhang H. TSVdb: a web-tool for TCGA splicing variants analysis. BMC Genomics; 2018; 1–7. https://bmcgenomics.biomedcentral.com/track/pdf/10.1186/s12864-018-4775-x

27. Silva TC, Colaprico A, Olsen C, Bontempi G, Ceccarelli M, Berman BP, et al. TCGAbiolinksGUI: A graphical user interface to analyze cancer molecular and clinical data [version 1; referees: 1 approved, 1 approved with reservations] Referee Status: 2018; https://doi.org/10.12688/f1000research.14197.1

28. Aran D, Butte AJ, Hanahan D, Coussens L, Aran D, Sirota M, et al. Digitally deconvolving the tumor microenvironment. Genome Biol. Genome Biology; 2016; 17: 175. https://doi.org/10.1186/s13059-016-1036-7 PMID: 27549319

29. Whiteside TL. The tumor microenvironment and its role in promoting tumor growth. Oncogene. 2008; 27: 5904–5912. https://doi.org/10.1038/onc.2008.271 PMID: 18836471

30. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. Nature Publishing Group; 2015; 6: 8971. https://doi.org/10.1038/ncomms9971 PMID: 26634437

31. Downing JR, Wilson RK, Zhang J, Mardis ER, Pui CH, Ding L, et al. The pediatric cancer genome project. Nat Genet. 2012; 44: 619–622. https://doi.org/10.1038/ng.2287 PMID: 22641210

32. Braakhuis BJM, Leemans CR, Brakenhoff RH. Using tissue adjacent to carcinoma as a normal control: An obvious but questionable practice. J Pathol. 2004; 203: 620–621. https://doi.org/10.1002/path.1549 PMID: 15141375

33. Frazee AC, Langmead B, Leek JT. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics. 2011; 12. https://doi.org/10.1186/1471-2105-12-449 PMID: 22087737

34. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nat Biotechnol. 2017; 35: 319–321. https://doi.org/10.1038/nbt.3838 PMID: 28398307

35. Wang Q, Armenia J, Zhang C, Penson A V., Reznik E, Zhang L, et al. Unifying cancer and normal RNA sequencing data from different sources. Sci Data. The Author(s); 2018; 5: 180061. https://doi.org/10.1038/sdata.2018.61 PMID: 29664468

36. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank. 2015; 13: 307–308. https://doi.org/10.1089/bio.2015.29031.hmm PMID: 26484569

37. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014; 15: R29. https://doi.org/10.1186/gb-2014-15-2-r29 PMID: 24485249

38. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009; 26: 139–140. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308

39. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28: 882–883. https://doi.org/10.1093/bioinformatics/bts034 PMID: 22257669

40. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. Nature Publishing Group; 2015; 12: 115–121. https://doi.org/10.1038/Nmeth.3252 PMID: 25633503

41. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. 2011;

42. Wickham H. Ggplot2. Wiley Interdiscip Rev Comput Stat. 2011; 3: 180–185. https://doi.org/10.1002/wics.147

43. Siu LL, Lawler M, Haussler D, Knoppers BM, Lewin J, Vis DJ, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. Nat Med. 2016; 22: 464–471. https://doi.org/10.1038/nm.4089 PMID: 27149219

44. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8: 118–127. https://doi.org/10.1093/biostatistics/kxj037 PMID: 16632515

45. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. Cell. 2018; 173: 338–354. e15. https://doi.org/10.1016/j.cell.2018.03.034 PMID: 29625051

46. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H, et al. The Immune Landscape of Cancer. Immunity. Cell Press; 2018; 48: 812–830.e14. https://doi.org/10.1016/J.IMMUNI.2018.03.023 PMID: 29628290

47. Espinoza JA, Jabeen S, Batra R, Papaleo E, Haakensen V, Timmermans Wielenga V, et al. Cytokine profiling of tumour interstitial fluid of the breast and its relationship with lymphocyte infiltration and clinicopathological characteristics. Oncoimmunology. 2016; 5: 00–00. https://doi.org/10.1080/2162402X.2016.1248015 PMID: 28123884

48. Terkelsen T, Haakensen VD, Saldova R, Gromov P, Papaleo E, Helland A, et al. N-glycan signatures identified in tumor interstitial fluid and serum of breast cancer patients: association with tumor biology and clinical outcome. Mol Oncol. 2018; 12: 972–990. https://doi.org/10.1002/1878-0261.12312 PMID: 29698574

49. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013; 4: 2612. https://doi.org/10.1038/ncomms3612 PMID: 24113773

50. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. Nature Publishing Group; 2012; 30: 413–421. https://doi.org/10.1038/nbt.2203 PMID: 22544022

51. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40: 4288–4297. https://doi.org/10.1093/nar/gks042 PMID: 22287627

52. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43: e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

53. Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, et al. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. Bioinformatics. 2016; btw575. https://doi.org/10.1093/bioinformatics/btw575 PMID: 27592709

54. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015; 163: 506–519. https://doi.org/10.1016/j.cell.2015.09.033 PMID: 26451490

55. Bušek P, Malík R, Šedo A. Dipeptidyl peptidase IV activity and/or structure homologues (DASH) and their substrates in cancer. Int J Biochem Cell Biol. 2004; 36: 408–421. https://doi.org/10.1016/S1357-2725(03)00262-0 PMID: 14687920

56. Collado-Torres L, Nellore A, Jaffe AE. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. F1000Research. 2017; 6: 1558. https://doi.org/10.12688/f1000research.12223.1 PMID: 29043067

57. Mochizuki S, Okada Y. ADAMs in cancer cell proliferation and progression. 2007; 98: 621–628. https://doi.org/10.1111/j.1349-7006.2007.00434.x PMID: 17355265

58. Berx G, van Roy F. Involvement of members of the cadherin superfamily in cancer. Cold Spring Harb Perspect Biol. 2009; 1. https://doi.org/10.1101/cshperspect.a003129 PMID: 20457567