Analyses of cancer data in the Genomic Data Commons Data Portal with new functionalities in the TCGAbiolinks R/Bioconductor package

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

[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]Department of Neurosurgery, Henry Ford Hospital, Detroit, MI, USA
[6]Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, FL 33136, USA

*corresponding authors: elenap@cancer.dk, axc1833@med.miami.edu

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 large data collection platform that includes different genomic studies included 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 without renouncing a robust statistical and theoretical framework are needed. 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 the Genotype-Tissue-Expression (GTEx) platform iv) support for other genomics datasets, here exemplified 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. Henceforth, we implemented these filtering procedures in TCGAbiolinks. Moreover, a limitation of some of the TCGA datasets is the unavailability or paucity of 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 v 2.8 and higher released in Bioconductor version 3.7.

Introduction

Cancer is among the leading causes of mortality worldwide, a complex disease where multiple different mechanisms are in play at the same time. The complexity of cancer lies in the fact that it is an 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 available a massive amount of data with 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, subsequently generating more than 2.5 petabytes of publicly available data over the past decade [8–10]. 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, and this was achieved through creating an "atlas" and applying large-scale genome-wide sequencing and multidimensional analyses. These latter 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 PanCancer 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 assembly. On the other hand, the harmonized version provides data that has been fully harmonized using GRCh38 (hg38) as a reference genome available through GDC portal.

Many tools have been so far developed to interface with the TCGA data [13–26] and 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 [17,18,27]. 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 such as fibroblast [28,29]. 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 [30], which the tools for analyses of TCGA data should employ.

Other cancer genomic initiatives have been following the TCGA model, such as the 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 [31]. Comprehensive support to the analyses of different genomic datasets with the same workflow is thus essential for both reproducibility and harmonization of the results.

At last, it is a 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 surrounding or adjacent to a highly genetically abnormal tumor is likely to show cancer-related molecular aberrations [32], biasing the comparison. Moreover, circulating biomolecules, originated 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* [33], *Recount2* [34] and *RNASEQDB* [35] where TCGA data were integrated with normal healthy samples from the Genotype-Tissue Expression (GTEx) project [36] have the potential to boost the comparative analyses also for those TCGA datasets where normal samples are underrepresented or unavailable.

We present new key features and enhancements that we implemented in *TCGAbiolinks* version 2.8 and higher in light of recent discoveries on the impact of quantification of the tumor purity of the samples under investigation [30], the need of a more substantial amount of normal samples [34], as well as the implementation of robust and statistically sound workflows for differential expression analyses [37,38] and exploration of potential sources of batch effects [39].

Design and Implementation

Overview on TCGAbiolinks

For the sake of clarity, we will at first briefly introduce the main functions of *TCGAbiolinks* that are extensively discussed in the original publication and a recently published workflow [17,18]. We advise referring directly to these publications and the vignette on *Bioconductor* for more details on the basic functionalities.

The data retrieval is handled by three main functions of *TCGAbiolinks*: *GDCquery*, *GDCdownload* and *GDCprepare* and allows 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 to interface with different genomics, transcriptomics and proteomics platforms, along with to retrieve clinical data, information on drug treatments, subtypes and biospecimen.

GDCprepare is especially relevant since it 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 [40] that is easily manageable and integrable with other *R/Bioconductor* packages or just as a dataframe for other forms of data manipulation.

Moreover, *TCGAbiolinks* offers the option to apply normalization methods with the function *TCGAanalyze_Normalization* adopting the *EDASeq* protocol [41], 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...) [42].

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

Toward more generalized analyses of genomic data in GDC

TCGAbiolinks was initially conceived to interact with the 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. As an example, we thus now worked to support the SE format also for other GDC datasets, such as the ones from the TARGET consortium which is included in*TCGAbiolinks 2.8*. The SE object provides the advantage of collecting clinical information on the samples (such as patient gender, age, 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 with the consequence of stalling the advancements in the analyses[43].The design of *GDCprepare* function of *TCGAbiolinks* thus nicely fulfils the need of standardized and harmonized ways to process data from different genomics initiatives which could find in GDC portal the common storage.

Handling batch corrections in TCGAbiolinks: TCGAbatch_correction

High-throughput sequencing or 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 factor 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 [39].

Here, we introduced the *TCGAbatch_correction* function to address and correct for different potential sources of batch effects linked to TCGA gene expression data using the *sva R* package [39]. The *sva* package provides a framework for removing artifacts either by (i) estimating surrogate variables that introduce unwanted variability in high-throughput high dimensional datasets or (ii) using the *ComBat* function that employ an empirical Bayesian framework to remove batch effects related to known sources [44]. Modeling for known batch effects significantly helps improving results by stabilizing error rates, 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 in which 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 [37] (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 verify 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 discussed in Figure 1).

TCGA_MolecularSubtype

Although each cancer is believed to be a single disease, advances in the genomic field pointed out that each cancer type is much more heterogeneous and 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 [17]. The format of the data is although not easy to navigate and integrate within other functions.

For this reason, we designed a new function *TCGA_MolecularSubtype* and of manually curated molecular subtypes for a total of 13 cancer types (Table 1). Collectively, we have molecular subtype annotation for 4768 individuals (of which 4469 with RNA-Seq data available). The function also allows fetching the subtype information not only for each cancer type, but also at for each TCGA barcode (i.e., for each individual sample).

In particular, the information used to classify cancer subtypes is the one used and the most recently published by the Pan-Cancer works from the TCGA consortium (http://bioinformaticsfmrp.github.io/TCGAbiolinks/subtypes.html#pancanceratlas_subtypes:_curated_molecular_subtypes). An alternative is also the function added in the context of the Pan-Cancer studies, namely *PanCancerAtlas_subtypes*, which provides molecular subtypes (in the column Subtype_Selected) for 24 cancer types and 7,734 TCGA's samples. These new functions have the advantage that the data are a curation retrieved from synapse directly and thus up-to-date (https://www.synapse.org/#!Synapse:syn8402849).

Recently we showed the advantage of using those functions to have a resource (in the same place) to quickly retrieve molecular subtypes in Pan-cancer studies and compares to novel defined stemness index [45] and immune subtypes [46] for individual TCGA's samples.

TCGAtumor_purity

The tumor microenvironment encloses cellular and non-cellular units that play a critical role in the initiation, progression, and metastasis of the tumor [30,47,48].

An important concept to retain from the TME definition is the 'tumor purity' which is defined 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. Currently, 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 [28].

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 sites [30]; iv) IHC, the Nationwide Children's Hospital Biospecimen Core Resource provided stain slides containing eosin and haemtoxylin which are processed using image analysis techniques to generate a tumor purity estimate [30]; v) Consensus measurement of Purity Estimation (CPE), a consensus estimate from the four methods mentioned above [30]. 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 by the *TCGAtumor_purity* function.

TCGAanalyze_DEA Extension

We revised and expanded the pre-existing *TCGAbiolinks* function that performed differential expression analyses (DEAs) calling the commonly used R package *edgeR*

[38]. In the former available 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). In particular 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 for case and control only thus 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 account for 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 to account for the patient factor in the case 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.

The function returns two types of objects: either i) a table with DEGs containing for each gene logFC, logCPM, p-value, and FDR corrected p-value in case of pairwise comparison, 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 coming from 18 published studies [33]. This atlas of RNA-Seq count table improves the process of data acquisition and allows cross-study comparisons since all the count tables 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 [53], and included in the recent R*ecount2* interface [34].

For this reason, *TCGAquery_Recount2* queries GTEx and TCGA *Recount2* for all tissues available on the online platform, providing the flexibility to the user to decide which tissue source to employ for the calculations.

TCGAquery_Recounts2 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*.

Examples

Below, we illustrate two cases studies as an example of the usage of the new functions and 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*.

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). Among 1222 BRCA samples available in GDC, for example purposes, 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 regarding both genes and samples and containing gene expression table of HTSeq-based counts from reads harmonized and aligned to hg38 genome assembly. Afterward, 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 fed to 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 understand the quality of the data and to identify possible anomalies or cofounder effects that need to be taken into account. This can be done 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, in this example, batch correction on TSS as a cofounder factor along with accounting for one covariate (cancer VS normal) and only two batches were retained. The results are reported in **Figure 1**.

According to the standard defined by the TCGA consortium, 60% purity is the recommended threshold for analyses [30]. Thus, we applied a filtering step where tumor samples that show a tumor purity less than 60% median CPE are discarded from the analysis using the *TCGAtumor_purity* function of *TCGAbiolinks*. 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* to exploit the power of generalized linear models beyond the control versus case scheme. As an illustrative case, we queried the

PAM50 classification [54] for each of the samples through *TCGA_MolecularSubtype* and then provided to the DEA method so the customizable *contrast.formula* argument can contain the formula designing the contrasts. Beforehand, data is normalized for GC-content, as explained above. As a final step, a quantile filtering is applied with a cutoff of 25%, as suggested by the original *TCGAbiolinks* workflow. Within the *TCGAanalyze_DEA* function, it is specified to also perform a *voom* transformation of the count data, as detailed above. In Figure 2, we show the results as a Volcano plot performing DEA with the new implementation of the *TCGAanalyze_DEA* function. For example, we identified the dipeptidyl peptidase-IV DPP6 as up-regulated in Luminal breast cancer subtypes with respect to normal samples. DPP6 belongs to a family of proteases that cleave X-Pro dipeptides from the N-terminal extremity of proteins. Several active peptides that have a role in cancerogenesis are enriched in conserved prolines as proteolytic-processing regulatory elements [55]. DPP6 overexpression could thus cause aberrant cellular functions. DPP-IV proteins have been also suggested as interesting therapeutic targets for developing inhibitors of their activity [55].

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, now it is possible to query data from the *Recount2* platform to increase the pool of normal samples and apply the DEA pipelines of T*CGAbiolinks*.

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 uterus tissues are used as a source of normal samples for DEA. Concerning the type of the queried count data, it is similarly harmonized HTSeq counts aligned to the genome assembly hg38 (see the script reported in https://github.com/ELELAB/TCGAbiolinks_examples). We used *TCGAquery_recount2* function to download tumour and normal uterus samples from the *Recount2* platform as Ranged Summarized Experiment (RSE) objects.

First, before engaging into 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 tackle before DEA. Hence, the user should extract sample metadata from RSE objects regarding read length and count of mapped reads 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 the equation (1) (possibly with the need to round the counts since the result might not be of an integer type).

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

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) [56]. 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 filtered with a 25% cut-off. The data were then fed to the *TCGAanalyze_DEA* function comparing normal versus cancer samples with the *limma-voom* pipeline. Two volcano plots that show the top down- and up-regulated genes are shown in Figure 3 and 4, respectively. As an example, we identified the up-regulated gene ADAM28 in UCS tumor samples when compared to the normal ones (logFC = 3.13). 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 promotion of cell growth and invasion [57]. We also identified other key players in cell adhesion such as the cadherin CDH1 [58] as top up-regulated genes in UCS .

Availability and Future Directions

The functions illustrated in this manuscript are now available in the 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 daily provide scientific advices to the github community within our https://github.com/BioinformaticsFMRP/TCGAbiolinks/issues to solve both software bugs and proving new functionalities needed/requested by the gGthub communities. The issues feature is a place where the users of *TCGAbiolinks* can share their experience with their analyses and case study that can be addressed by our team or other Github users as well.

The newly developed functions will allow for the first time to fully appreciate the effect of using genuinely healthy samples or normal tumor-adjacent samples as a control as well as to correct for tumor purity of the samples. We provided a more robust and comprehensive workflow to carry out differential expression analysis with two different methods and a customizable design matrix, as well as capability to handle batch corrections, which overall will provide the community with the possibility to use the same framework for example for benchmarking of differential expression methods (https://bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/extension.html).

Acknowledgments

The project was supported by a KBVU Pre-graduate scholarship 2017 to M.L. in EP group, as well as the LEO foundation grant number LF17006 and Innovation Fund Denmark grant number 5189-00052B. This work has also been supported by the

BridgeIRIS project [http://mlg.ulb.ac.be/BridgeIRIS], funded by INNOVIRIS, Region de Bruxelles Capitale, Brussels, Belgium, and by GENGISCAN: GENomic profiling of Gastrointestinal Inflammatory-Sensitive CANcers, [http://mlg.ulb.ac.be/GENGISCAN] Belgian FNRS PDR (T100914F to AC, CO and GB) and by a grant from Henry Ford Hospital (HN) and by the São Paulo Research Foundation (FAPESP) (2016/01389-7 to TCS & HN and 2015/07925-5 to HN). The authors would like to thank Matteo Tiberti for fruitful discussion and comments.

Author contributions

Conceptualization: EP; Data curation: MM,ML,TCS,AC,EP; Formal Analysis: MM, EP, ML; Funding Acquisition: EP; Investigation: MM, ML, EP, AC; Methodology: MM, ML, EP, AC; Project Administration: EP; Resources: EP; Software: MM, TCS, AC; Supervision: EP, AC; Validation: EP, ML; Visualization: MM; Writing-Original Draft Preparation: MM, EP; Writing-Review and Editing: MM, ML, TCS, HN, GB, AC, CO.

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. doi:10.1038/nrc3261
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. doi:10.1038/nature12625
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63. doi:10.1038/nrg2484
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. doi:10.1038/onc.2015.90
5. Van Verk MC, Hickman R, Pieterse CMJ, Van Wees SCM. RNA-Seq: Revelation of the messengers. Trends Plant Sci. 2013;18: 175–179. doi:10.1016/j.tplants.2013.02.001
6. McGettigan PA. Transcriptomics in the RNA-seq era. Curr Opin Chem Biol. 2013;17: 4–11. doi:10.1016/j.cbpa.2012.12.008
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. doi:10.3390/cancers7030869
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. doi:10.1038/ng.2764
9. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Wspolczesna Onkol. 2015;1A: A68–A77. doi:10.5114/wo.2014.47136
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. doi:10.3389/fcell.2017.00083

11. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell. Elsevier Inc.; 2018;173: 283–285. doi:10.1016/j.cell.2018.03.042
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. doi:http://dx.doi.org/10.1109/MCSE.2016.92
13. Samur MK. RTCGAToolbox: A New Tool for Exporting TCGA firehose data. PLoS One. 2014;9. doi:10.1371/journal.pone.0106397
14. João F Matias Rodrigues1, Thomas SB Schmidt1 JT and C von M. TCGA-Assembler 2: Software Pipeline for Re- trieval and Processing of TCGA/CPTAC Data. Bioinformatics. 2017; 0–0. doi:10.1093/bioinformatics/xxxxx
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. doi:10.1371/journal.pone.0165395
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. doi:10.1038/srep02652
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-. doi:10.1093/nar/gkv1507
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. doi:10.12688/f1000research.8923.1
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. doi:10.1093/nar/gkx247
20. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. PeerJ Comput Sci. 2016;2: e67. doi: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. doi:10.1093/nar/gkv1478
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. doi:10.1186/s12859-016-0917-9
23. Wan Y-W, Allen GI, Liu Z. TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R. Bioinformatics. 2015; btv677-. doi:10.1093/bioinformatics/btv677
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. doi:10.1093/nar/gkv1288
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. doi:10.1093/bib/bby023
26. Zhang H. TSVdb: a web-tool for TCGA splicing variants analysis. BMC Genomics; 2018; 1–7. Available: 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; doi: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. doi:10.1186/s13059-016-1036-7
29. Whiteside TL. The tumor microenvironment and its role in promoting tumor growth. Oncogene. 2008;27: 5904–5912. doi:10.1038/onc.2008.271
30. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. Nature Publishing Group; 2015;6: 8971. doi:10.1038/ncomms9971
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. doi:10.1038/ng.2287
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. doi:10.1002/path.1549
33. Frazee AC, Langmead B, Leek JT. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics. 2011;12. doi:10.1186/1471-2105-12-449
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. doi:10.1038/nbt.3838
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. doi:10.1038/sdata.2018.61
36. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank. 2015;13: 307–308. doi:10.1089/bio.2015.29031.hmm
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. doi:10.1186/gb-2014-15-2-r29
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. doi:10.1093/bioinformatics/btp616
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. doi:10.1093/bioinformatics/bts034
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. doi:10.1038/Nmeth.3252
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. doi: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. doi:10.1038/nm.4089
44. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8: 118–127. doi:10.1093/biostatistics/kxj037
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. doi:10.1016/j.cell.2018.03.034

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. doi:10.1016/J.IMMUNI.2018.03.023

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. doi:10.1080/2162402X.2016.1248015

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. doi:10.1002/1878-0261.12312

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. doi:10.1038/ncomms3612

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. doi:10.1038/nbt.2203

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. doi:10.1093/nar/gks042

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. doi:10.1093/nar/gkv007

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. doi:10.1093/bioinformatics/btw575

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. doi:10.1016/j.cell.2015.09.033

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. doi:10.1016/S1357-2725(03)00262-0

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

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

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

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

TCGA study	Number of samples with both subtype information and RNASEQ data available
TCGA-BRCA	942
TCGA-GBM	156
TCGA-LUAD	290
TCGA-UCEC	391
TCGA-KIRC	513
TCGA-HNSC	314
TCGA-LGG	511
TCGA-THCA	549
TCGA-LUSC	193
TCGA-PRAD	377
TCGA-SKCM	66
TCGA-KICH	89
TCGA-ACC	78

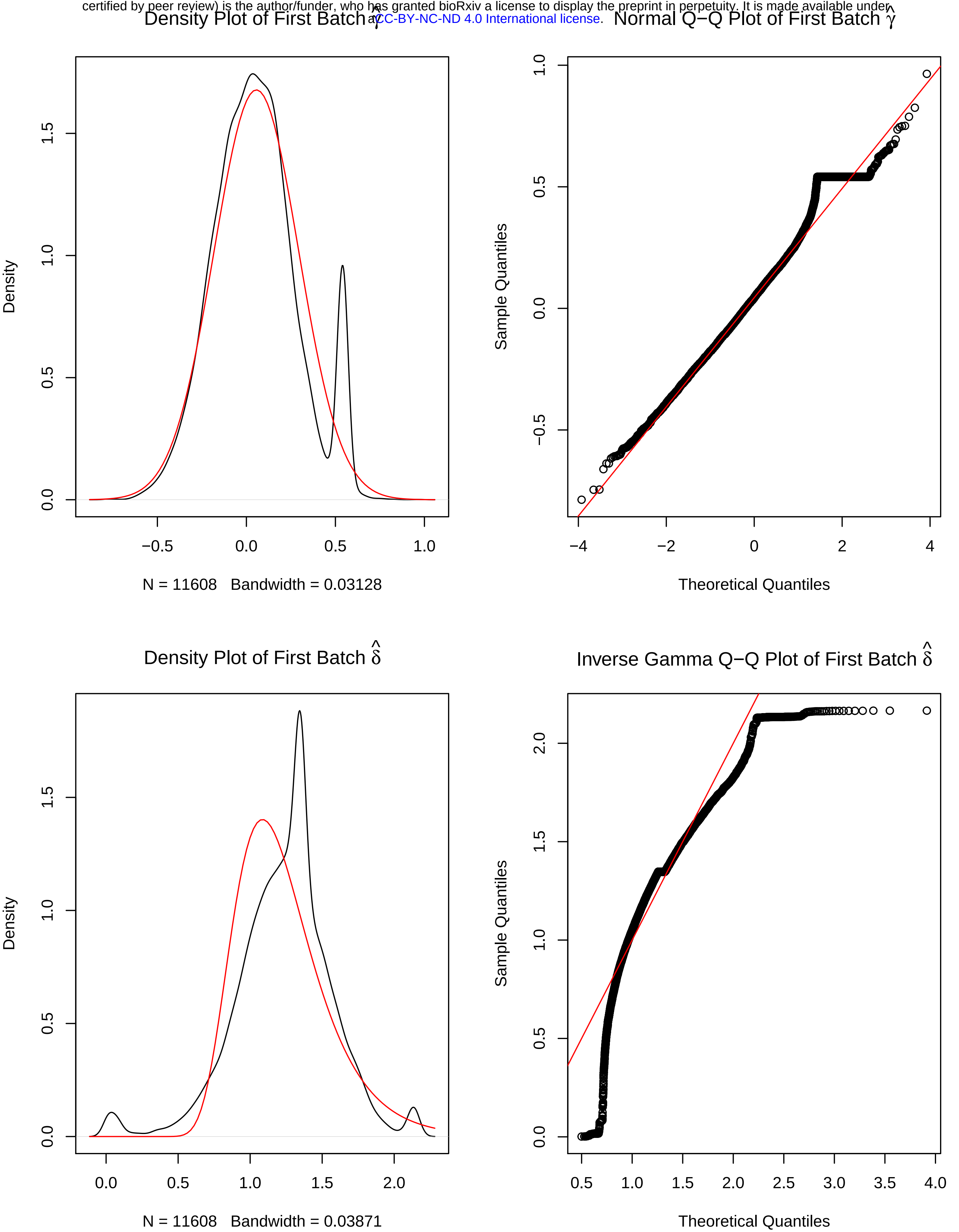
Figure Legends

Figure 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 only to avoid batches containing only one sample.

Figure 2. DEA analyses of TCGA-BRCA data comparing luminal subtypes with normal samples. A volcano plot is shown where only those genes with logFC higher than 5 or lower than -5 are shown as representative up- and down-regulated genes, respectively when comparing the Luminal subtype to the normal breast samples.

Figure 3. Down-regulated genes in uterine cancer compared to healthy uterus tissue samples. In the volcano plot, the down-regulated genes with logFC lower than -5 are shown as a result of DEA carried out comparing primary tumor samples from TCGA-UCS and normal uterus tissue samples from GTEx.

Figure 4. Up-regulated genes in uterine cancer compared to healthy uterus tissue samples. In the volcano plot, the up-regulated genes with logFC higher than 5 are shown as a result of DEA carried out comparing primary tumor samples from TCGA-UCS and normal uterus tissue samples from GTEx.

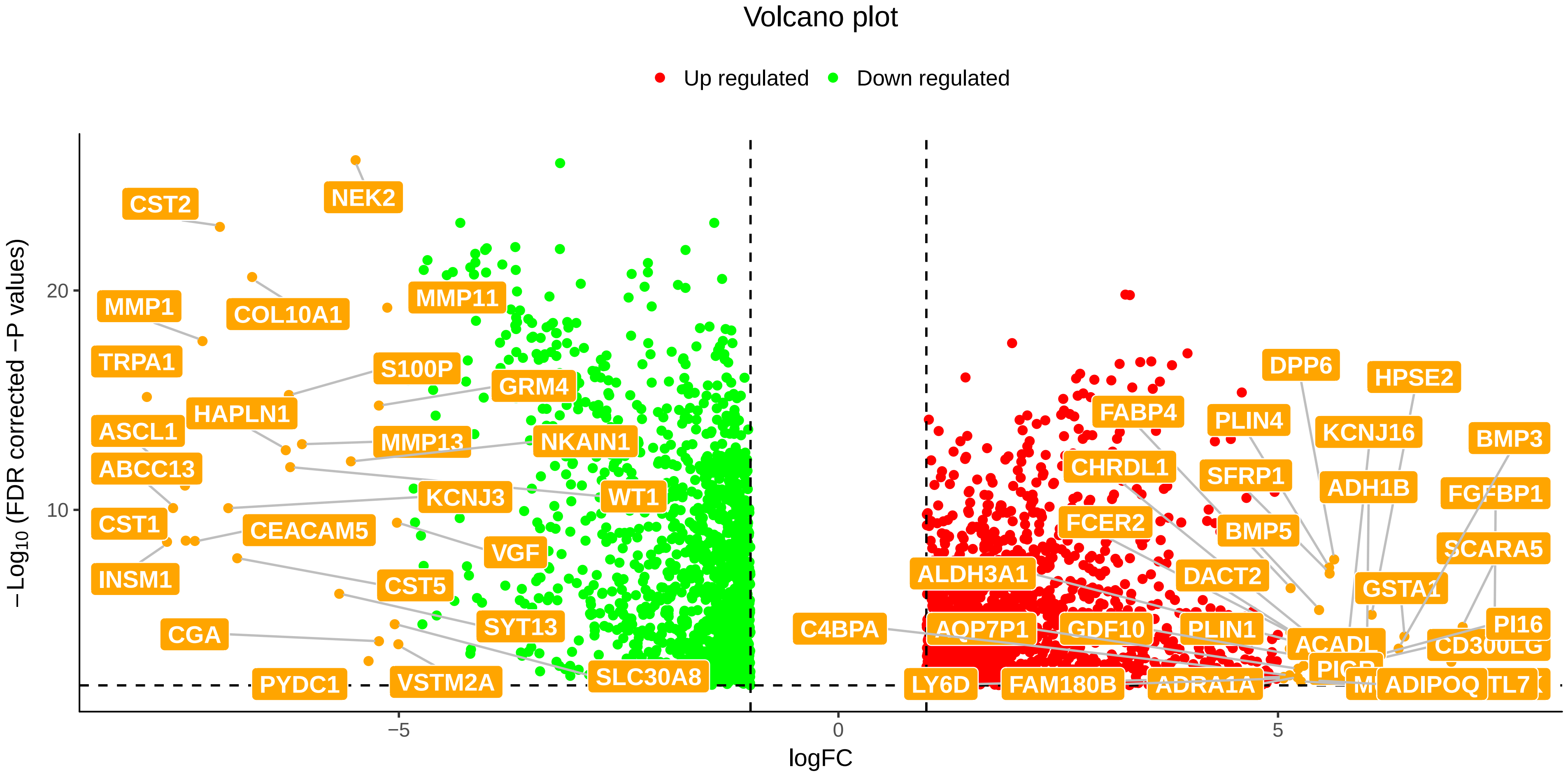


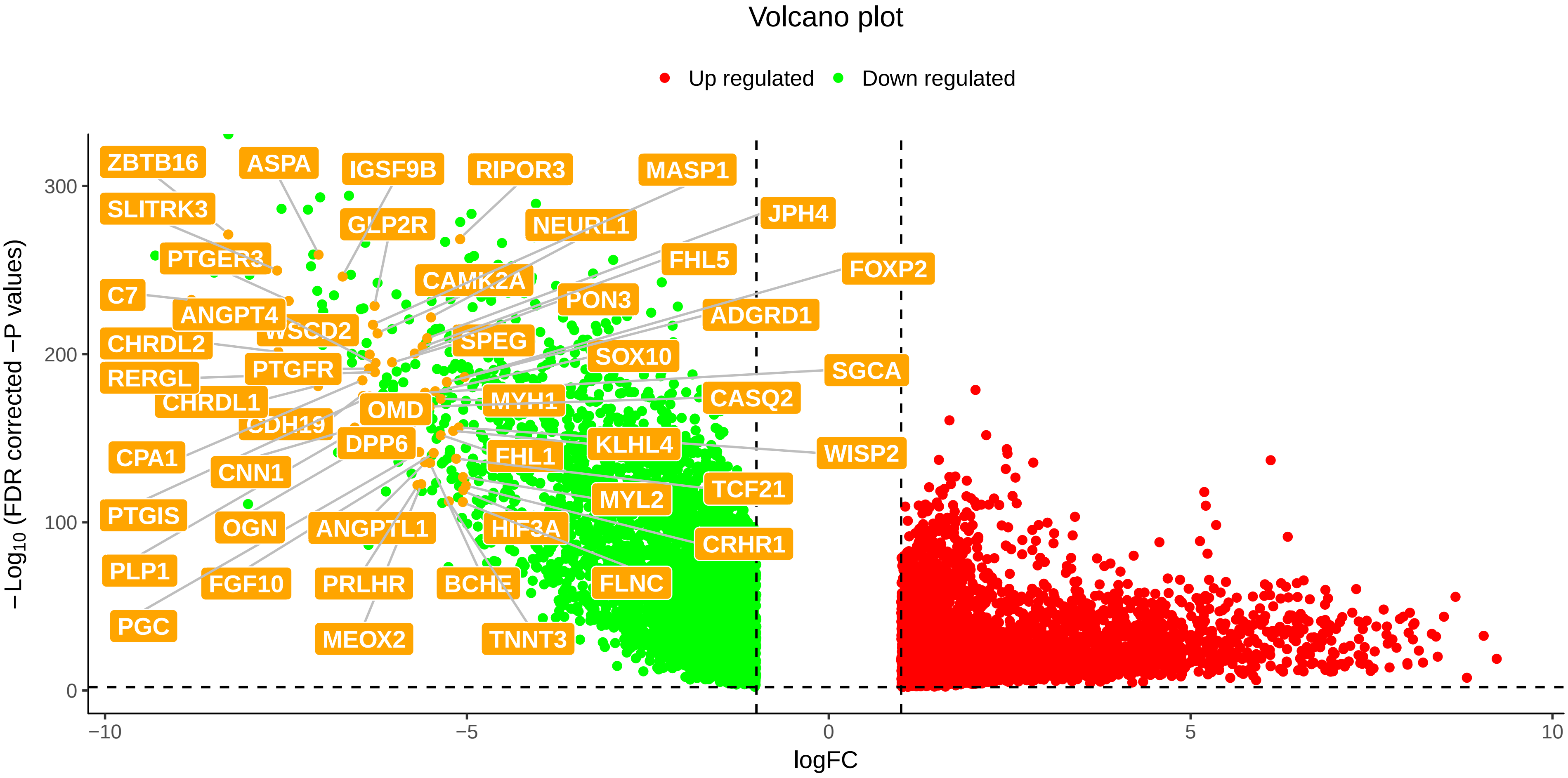
Density Plot of First Batch $\hat{\gamma}$

Normal Q−Q Plot of First Batch $\hat{\gamma}$

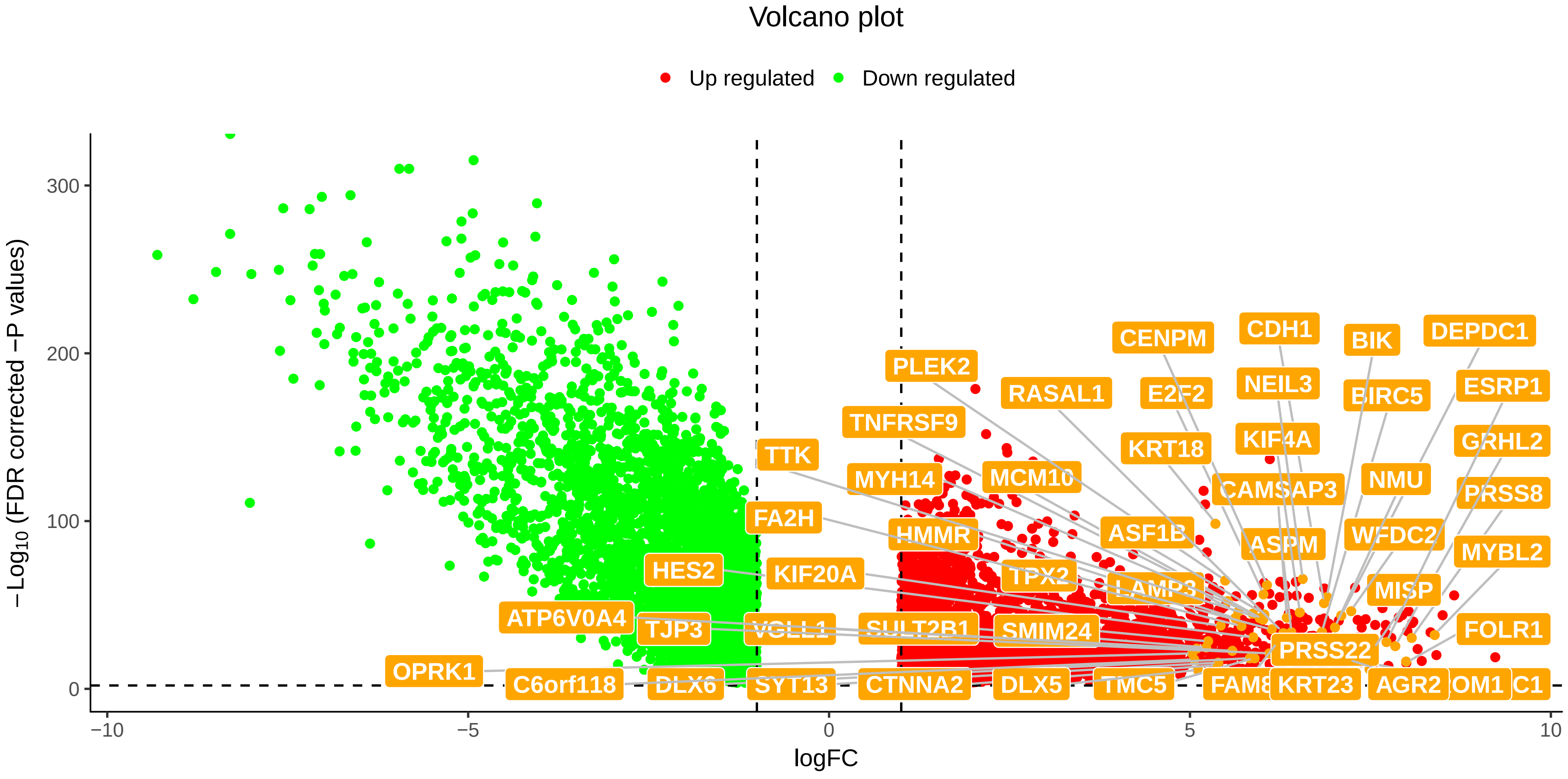
Density Plot of First Batch $\hat{\delta}$

Inverse Gamma Q−Q Plot of First Batch $\hat{\delta}$


Volcano plot
● Up regulated ● Down regulated
$-\mathrm{Log}_{10}$ (FDR corrected −P values)
20
10
−5
0
5
logFC
NEK2
CST2
MMP1
COL10A1
MMP11
TRPA1
S100P
GRM4
HAPLN1
ASCL1
MMP13
NKAIN1
ABCC13
KCNJ3
WT1
CST1
CEACAM5
VGF
INSM1
CST5
SYT13
CGA
PYDC1
VSTM2A
SLC30A8
C4BPA
LY6D
ALDH3A1
AQP7P1
GDF10
FAM180B
ADRA1A
PLIN1
FABP4
CHRDL1
FCER2
DACT2
PLIN4
SFRP1
BMP5
DPP6
HPSE2
KCNJ16
ADH1B
GSTA1
ACADL
PIGR
M
ADIPOQ
TL7
BMP3
FGFBP1
SCARA5
PI16
CD300LG



Volcano plot
● Up regulated
● Down regulated
ZBTB16
ASPA
IGSF9B
RIPOR3
MASP1
SLITRK3
GLP2R
NEURL1
JPH4
PTGER3
FHL5
CAMK2A
FOXP2
C7
PON3
ANGPT4
ADGRD1
WSCD2
CHRDL2
SPEG
SOX10
RERGL
PTGFR
SGCA
CHRDL1
MYH1
CASQ2
OMD
CDH19
DPP6
KLHL4
WISP2
CPA1
FHL1
CNN1
TCF21
MYL2
PTGIS
OGN
ANGPTL1
HIF3A
CRHR1
PLP1
FGF10
PRLHR
BCHE
FLNC
PGC
MEOX2
TNNT3
300
200
100
0
−Log$_{10}$ (FDR corrected −P values)
−10
−5
0
5
10
logFC


Volcano plot
● Up regulated
● Down regulated
−Log10 (FDR corrected −P values)
300
200
100
0
logFC
−10
−5
0
5
10
CENPM
CDH1
BIK
DEPDC1
PLEK2
RASAL1
E2F2
NEIL3
BIRC5
ESRP1
TNFRSF9
KRT18
KIF4A
GRHL2
TTK
MYH14
MCM10
CAMSAP3
NMU
PRSS8
FA2H
HMMR
ASF1B
ASPM
WFDC2
MYBL2
HES2
KIF20A
TPX2
LAMP3
MISP
ATP6V0A4
TJP3
VGLL1
SULT2B1
SMIM24
FOLR1
PRSS22
OPRK1
C6orf118
DLX6
SYT13
CTNNA2
DLX5
TMC5
FAM8
KRT23
AGR2
OM1
C1