frontiers in Genetics

TECHNOLOGY AND CODE
published: 18 February 2021
doi: 10.3389/fgene.2021.624259

Galaxy and MEAN Stack to Create a User-Friendly Workflow for the Rational Optimization of Cancer Chemotherapy

Jorge Guerra Pires[1†], Gilberto Ferreira da Silva[1†], Thomas Weyssow[2], Alessandra Jordano Conforte[1,3], Dante Pagnoncelli[4], Fabricio Alves Barbosa da Silva[3] and Nicolas Carels[1†]*

[1] Plataforma de Modelagem de Sistemas Biológicos, Center for Technology Development in Health (CDTS), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, [2] Informatic Department, Free University of Brussels (ULB), Brussels, Belgium, [3] Laboratório de Modelagem Computacional de Sistemas Biológicos, Scientific Computing Program, FIOCRUZ, Rio de Janeiro, Brazil, [4] Instituto COI, Rio de Janeiro, Brazil

OPEN ACCESS

Edited by:
Fatemeh Maghuly, University of Natural Resources and Life Sciences, Vienna, Austria

Reviewed by:
Julie Krainer, Austrian Institute of Technology (AIT), Austria
Vishal Sarsani, University of Massachusetts Amherst, United States

***Correspondence:**
Nicolas Carels
nicolas.carels@cdts.fiocruz.br; nicolas.carels@gmail.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *31 October 2020*
Accepted: *22 January 2021*
Published: *18 February 2021*

Citation:
Pires JG, Silva GF, Weyssow T, Conforte AJ, Pagnoncelli D, Silva FAB and Carels N (2021) Galaxy and MEAN Stack to Create a User-Friendly Workflow for the Rational Optimization of Cancer Chemotherapy. Front. Genet. 12:624259. doi: 10.3389/fgene.2021.624259

One aspect of personalized medicine is aiming at identifying specific targets for therapy considering the gene expression profile of each patient individually. The real-world implementation of this approach is better achieved by user-friendly bioinformatics systems for healthcare professionals. In this report, we present an online platform that endows users with an interface designed using MEAN stack supported by a Galaxy pipeline. This pipeline targets connection *hubs* in the subnetworks formed by the interactions between the proteins of genes that are up-regulated in tumors. This strategy has been proved to be suitable for the inhibition of tumor growth and metastasis *in vitro*. Therefore, Perl and Python scripts were enclosed in Galaxy for translating RNA-seq data into protein targets suitable for the chemotherapy of solid tumors. Consequently, we validated the process of target diagnosis by (i) reference to subnetwork entropy, (ii) the critical value of density probability of differential gene expression, and (iii) the inhibition of the most relevant targets according to TCGA and GDC data. Finally, the most relevant targets identified by the pipeline are stored in MongoDB and can be accessed through the aforementioned internet portal designed to be compatible with mobile or small devices through Angular libraries.

Keywords: systems biology, translational oncology, personalized medicine, Galaxy, MEAN stack, angular, protein–protein network, Shannon entropy

INTRODUCTION

The worldwide estimate of people diagnosed with cancer was 18.1 million in 2017[1] and it is predicted by the *World Health Organization* (WHO) to be 27 million new cases worldwide by 2030. On its own, breast cancer (BC) continues to be among the most frequent cancer around the world alongside the prostate one. Moreover, BC, alone accounts for almost 2.1 million new cases diagnosed annually worldwide, causing an estimate of 600,000 deaths every year (Bray et al., 2018). Because of these dire statistics, BC has received huge attention from both the academic and the

[1] https://ourworldindata.org/cancer

industry, which resulted in a large corpus of publication (culminating at 25,000 in 2019[2]) and publicly available datasets.

In addition, the well-known heterogeneity of breast cancer has justified the genomic study of tumors on a large scale in search for tumor subtypes that could allow a better understanding of the tumor biology and could serve as support for the establishment of genetic signatures, which, when validated in clinical trials, could pave the way for an increasingly specific and more precise treatment than the clinical parameters currently in use.

It is a more in-depth knowledge of tumor biology that has allowed for greater individualization of available treatments and has made it possible to overcome the relapse and resistance eventually observed with traditional treatments (Naito and Urasaki, 2018). In addition, clinical experience has shown that knowledge of the individual characteristics of each tumor may contribute to better therapeutic results with less toxicity.

According to the *one-size-fits-all* approach of chemotherapy, treatment should fit every individual of a population. As a consequence, it is intrinsically imprecise since it does not take into account the genetic peculiarities of each patient. Thus, a one-size-fits-all treatment approach does not work for everyone and may cause harmful side effects. By contrast, *personalized oncology*, which can be placed into a wider paradigm shift called *personalized medicine*, involves the tailoring of medical treatment to the individual characteristics or symptoms and responses of a patient during all stages of care.

The paradigm of one-size-fits-all treatment is now undergoing a shift toward personalized oncology with the identification of molecular pathways predicting both tumor biology as well as response to therapy. Most of those achievements have been inserted into mathematical and computational models by different groups, which can be used to test therapies and hypothesis; the one presented herein fall into this category.

A *new taxonomy* of disease based on molecular and environmental determinants rather than signs and symptoms has been proposed (Collins and Varmus, 2015). The paradigm revolution lies in the change from a clinician selecting a generic therapy on a heuristic basis to one based on molecular facts, a process called *evidence-based medicine* (Masic et al., 2008).

The tools of systems biology made it possible to analyze the huge amount of data delivered by high throughput technologies (broadly named Big Data, Willems et al., 2019). At the moment, the most common strategy for implementing high throughput technologies in oncology is to map mutations that promote suppressor and oncogenes (Guo et al., 2014; Campbell et al., 2020), which is a typical activity of *pharmacogenomics*. Briefly, pharmacogenomics aims at understanding why individuals respond differently to medicines on a genetic level. Consequently, it enables one to predict an individual's response to a drug according to genetic information and allows one to choose the most appropriate medication according to an individual's genetic composition. Furthermore, when the molecular diagnosis is performed, targeted therapy is designed for acting on specific molecular targets supposed to be relevant for the tumor under consideration (Wilsdon et al., 2018). Notwithstanding all the knowledge we have gathered so far, the relevance of a drug target is not obvious, and many criteria were pursuit in that quest (Catharina et al., 2018).

The development of personalized medicine is directly related to the availability of high-throughput technologies. High-throughput techniques, such as microarray, *RNA sequencing* (RNA-seq), and nanoString[3] are important tools for the characterization of tumors and their adjacent non-malignant tissues (Finak et al., 2006). Therefore, these techniques allow a better understanding of tumor biology (Carels et al., 2020). In particular, RNA-seq analysis through in silico methodologies demonstrated that each tumor is unique considering the protein profile of their up-regulated genes (Carels et al., 2015a).

Following the current state of the art, there are mainly two types of omics tests: (i) prognostic tests, which predicts a clinical outcome, and (ii) therapy guiding tests (theranostics), which enable the identification of patient subgroups with a similar response to a particular therapy (McShane and Polley, 2013). In this report, we focus on theranostics.

A variety of multigene assays are in clinical use or under investigation, which further defines the molecular characteristics of the cancers' dominant biologic pathways. Even if there has been a growing use of biomarkers in clinical trials, the use of single-marker and panel tests is still limited (Vuckovic et al., 2016). Gaining insight into the molecular composition of each tumor is recommended for eliminating the misuse of ineffective and potentially harmful drugs.

Mapping gene alterations by reference to the genome is generally performed to characterize indirect relationships between tumor development and indels, mutations, hyper- or hypo-methylation. By contrast, the description of transcriptome, proteome, or metabolome allows the characterization of a molecular phenotype. Interestingly, most *companion diagnostics* (CD) for cancer characterization on the market are based on mutation profiling. Accordingly, CDs are expected to guide the application of a specific therapy supposed to be efficacious for a given patient's condition (Verma, 2012). As a result, CDs allow the selection of a treatment that is more likely to be effective for each individual based on the genetic signatures of their tumors. Moreover, CDs are also developed for better predicting the patient response to a given treatment.

An approach based on molecular phenotyping recently proposed was the identification of the most relevant protein targets for specific therapeutic intervention in malignant BC cell lines (Carels et al., 2015a) based on the diagnosis of up-regulated interactome hubs. This strategy combined *protein-protein interactions* (PPI) and RNA-seq data for inferring (i) the topology of the signaling network of up-regulated genes in malignant cell lines and (ii) the most relevant protein targets therein. Hence, it has the benefit to allow the association of a drug to the entropy of a target and, additionally, to rank drugs according to their respective entropy by reference to their targets (Carels et al., 2015b).

[2]https://pubmed.ncbi.nlm.nih.gov/?term=breast+cancer

[3]https://www.nanostring.com

Three concepts were considered in the approach followed by Carels et al. (2015a): (i) A vertex with a high expression level is more influential than a vertex with a low expression level. (ii) A vertex with a high connectivity level (hub) is more influential than a vertex with a low connectivity level. (iii) A protein target must be expressed at a significantly higher level in tumor cells than in the cells used as a non-malignant reference to reduce harmful side effects to the patient after its inhibition. It is worth mentioning that each combination of targets that most closely satisfied these conditions was found to be specific for its respective malignant cell lines. These statements were validated *in vitro* on a BC model by Tilli et al. (2016). These authors showed that the inactivation, by *small interfering RNA* (siRNA), of the five top-ranked hubs of connection (top-5) identified for MDA-MB-231, a triple-negative cell line of invasive BC, resulted in a significant reduction of cell proliferation, colony formation, cell growth, cell migration, and cell invasion. Inhibition of these targets in other cell lines, such as MCF-7 (non-invasive malignant breast cell line) and MCF-10A (non-tumoral cell line used as a control), showed little or no effect, respectively. In addition, the effect of joint target inhibition was greater than the one expected from the sum of individual target inhibitions, which is in line with the buffer effect of regulatory pathway redundancy in malignant cells (Tilli et al., 2016).

The signaling network of a biological system is scale-free (Albert et al., 2000), which means that few proteins have high connectivity values and many proteins have low connectivity values. As proven mathematically, the inhibition of proteins with high connectivity values has a greater potential for signaling network disruption than randomly selected proteins (Albert et al., 2000). This evidence was proven in silico by Conforte et al. (2019) in the particular case of tumor signaling networks.

In terms of systems biology, the inhibitory activity of a drug may be modeled by the removal of its corresponding protein target from the signaling network to which it belongs (Carels et al., 2015b; Conforte et al., 2019). The impact of vertex removal from a network can be evaluated by the use of the Shannon entropy, which has been proposed as a network complexity measure and applied by many authors to determine a relationship between network entropy and tumor aggressiveness. Breitkreutz et al. (2012), for instance, inferred a negative correlation between the entropy of networks made of genes documented in the *Kyoto Encyclopedia of Genes and Genomes* (KEGG[4]) database considering cancer types and their respective 5-year survival. The existence of this negative correlation was demonstrated later on by Conforte et al. (2019) using RNA-seq data from bench experiments stored in *The Cancer Genome Atlas* (TCGA now hosted by the *Genomic Data Commons Data Portal* – GDC Data Portal[5]).

The Shannon entropy (H) is given by formula 1

$$H = -\sum_{k=1}^{n} p\left(k\right) log_2\left(p\left(k\right)\right) \quad (1)$$

where $p(k)$ is the probability that a vertex with a connectivity value k occurs in the analyzed network.

The process of multistep mining of high throughput data can be cumbersome to handle by humans and needs translation into machine language and automation (Deelman et al., 2009). Thus, according to the scientific challenge, we developed codes in Perl and Python. To deal with assembling a workflow based on *heterogeneous programming*, i.e., a workflow including more than one programming language, we chose Galaxy (Afgan et al., 2018) that fit this purpose.

Since we believe that a molecular phenotyping strategy is worthwhile for complementing the genotyping approach, we described in this report how to perform the translation from RNA-seq data into therapy targets based on the process described in more detail in Conforte et al. (2019). The most relevant targets stored in MongoDB can be accessed through an internet portal written in JavaScript using the software bundle called MEAN stack and portable to mobile and small devices through Angular Flex-Layout library and *Lazy loading*[6] strategies as described by Fain and Moiseev (2018) and Holmes and Herber (2019).

MATERIALS AND METHODS

Galaxy Pipeline

TCGA Data

The gene expression data were obtained as RNA-seq files from paired samples (control and tumor samples from the same patient) and downloaded from TCGA[7] in February 2016 and from the GDC Data Portal[8] in March 2020. The data selection followed two criteria: (i) for each cancer type, approximately 30 patients with paired samples were required to satisfy statistical significance; and (ii) the tumor samples had to be from a solid tumor. The data from TCGA and GDC are given in **Table 1**.

In TCGA, gene expression values were given for 20,532 genes referred to as GeneSymbol, calculated by *RNA-seq through*

[4]http://www.genome.jp/kegg

[5]https://portal.gdc.cancer.gov

[6]https://en.wikipedia.org/wiki/Lazy_loading. Accessed on 14/10/2020.

[7]https://cancergenome.nih.gov/

[8]https://portal.gdc.cancer.gov/

TABLE 1 | RSEM-UQ from paired tumor-stroma data retrieved from TCGA and FPKM-UQ from GDC.

Tumor type	Abbreviation	OS[1]	TCGA, n[2]	GDC, n
Stomach adenocarcinoma	STAD	38	32	27
Lung adenocarcinoma	LUAD	40	57	57
Lung squamous cell carcinoma	LUSC	47	50	48
Liver hepatocellular carcinoma	LIHC	49	49	50
Kidney renal clear cell carcinoma	KIRC	63	71	71
Kidney renal papillary cell carcinoma	KIRP	75	32	31
Breast cancer	BRCA	82	72	46
Thyroid cancer	THCA	93	57	56
Prostate cancer	PRAD	98	51	50

[1]OS: 5-years overall survival taken from Liu et al. (2018) according to Conforte et al. (2019), %. [2]n: Sample size, number.

expectation maximization (RSEM) (Mortazavi et al., 2008; Li and Dewey, 2011). Since they were normalized according to the upper quartile methods (formula 2) as reported in GDC documentation[9], we denoted them as RSEM-UQ. In the case of GDC, gene expression values were given for 60,483 sequences, calculated by FPKM and referred to as Ensembl accession number. As those values were also normalized by upper quartile, they were denoted, here, as FPKM-UQ. We considered RNA-seq from BRCA and LUAD as non-significant because of inconsistencies between *raw counts* file names, which led to a final sample of 16 and 17 for LUAD and BRCA, respectively. The 14,126 genes for which the equivalence between GeneSymbols and UniProtKB could be obtained went through further analysis.

$$N_{norm} = \frac{RC_g * 10^9}{RC_{g75} * L} \quad (2)$$

where:

RC_g: Number of reads mapped to the gene;

RC_{g75}: The 75th percentile read count value for genes in the sample;

L: Length of the coding sequence in base pairs.

ArrayEXPRESS Data

Fastq files from RNA-seq of tumor-stroma paired samples from 14 PRAD[10], and 18 *non-small cell lung cancer* (NSCLC)[11], were retrieved from ArrayEXPRESS[12]. These files were compared to the proteins of the EBI's interactome (see below) using BLASTx and processed through our pipeline to measure the average entropies of malignant up-regulated genes from both PRAD and NSCLC. The statistical significance of average entropy differences between PRAD and NSCLC was assessed through the Student's *t*-test using formula 3:

$$u_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{SCE_1}{n_1(n_1-1)} + \frac{SCE_2}{n_2(n_2-1)}}} \quad (3)$$

where:

$\bar{x}_i$: The average of sample i;

SCE_i: the sum of squared differences of sample i;

n_i: the size of sample i.

Because sample sizes of PRAD (n = 14) and NSCLC (n = 18) were less than n = 20, u_{obs} was compared to the theoretical value $t_{1-\alpha/2}$ of the Student's distribution using the k degree of freedom calculated according to formula 4 (Welch, 1949; Dagnelie, 1970):

$$k = \frac{\left[\frac{SCE_1}{n_1(n_1-1)} + \frac{SCE_2}{n_2(n_2-1)}\right]^2}{\frac{1}{n_1-1}\left[\frac{SCE_1}{n_1(n_1-1)}\right]^2 + \frac{1}{n_2-1}\left[\frac{SCE_2}{n_2(n_2-1)}\right]^2} \quad (4)$$

with n_1-1 < k < $n_1 + n_2$-2.

[9]https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/

[10]https://www.ebi.ac.uk/ena/data/view/PRJEB2449

[11]https://www.ebi.ac.uk/ena/data/view/PRJNA320473

[12]https://www.ebi.ac.uk/arrayexpress/

Identification of Hubs Among Genes Up-Regulated in Tumor Samples

To identify genes that were significantly differentially expressed in the tumor samples of patients, we subtracted gene expression values of control samples from their respective tumor paired samples. The resulting values were called differential gene expression. Negative differential gene expression values indicated higher gene expressions in control samples, while positive differential gene expression values indicated higher gene expressions in tumor samples.

The histogram of differential expression was normalized with the Python packages *scipy*. We used the probability density and cumulative distribution functions, respectively abbreviated as PDF and CDF, in the interval of differential gene expression from −20.000 to +20.000, to calculate the critical value corresponding to the one-tail cumulated probability p = 0.975, which corresponded to a p-value α = 0.025. We considered the genes as up-regulated when their differential expression was larger than the critical value corresponding to p = 0.975. The −20.000 to +20.000 range worked fine for the p-value and normalization conditions presented in this report. However, some normalization procedures flatten the probability distribution with Bayesian functions for variance minimization. Under these conditions, a p–value of 0.001 may represent a very large critical value of 80,000 or more, which would induce the scipy package to return "out of range." To beat this challenge, we introduced the possibility of tuning the −20.000 to +20.000 range to allow the user to try other normalization conditions together with more restrictive p-values. However, for coherence, all the data produced in this report were obtained with critical values in the −20.000 to +20.000 range.

In a subsequent step, the protein–protein interaction (PPI) subnetworks were inferred for the proteins identified as products of up-regulated genes. The subnetworks were obtained by comparing these gene lists with the human interactome.

The human interactome (151,631 interactions among 15,526 human proteins with UniProtKB accessions) was obtained from the intact-micluster.txt file (version updated December 2017) accessed on January 11, 2018[13].

We used the PPI subnetworks of up-regulated genes from each patient to identify each vertex (protein) degree through automated counting of their edges. These values were used to calculate the Shannon entropy of each PPI subnetwork as explained in the section "Shannon Entropy" below.

Shannon Entropy

The Shannon entropy was calculated with formula 1, where $p(k)$ is the probability of occurrence of a vertex with a rank order k (k edges) in the subnetwork considered. The subnetworks were generated automatically from gene lists found to be up-regulated in each patient.

Validation Process

The diagnosis of up-regulated genes with a higher vertex degree, which we considered as the most relevant target here, depends

[13]ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact-micluster.txt

on how *fastq* and *raw count* files are processed. First, *fastq* reads need to be transformed into *raw counts* and, second, *raw counts* need to be normalized. For validating this process, we used the data of RSEM-UQ from TCGA as available in 2016 that we referenced to as TCGA RSEM-UQ below. When referring to the FPKM-UQ files from GDC accessed in March 2020, we denoted them as *GDC FPKM-UQ*. Since we had no access to the raw counts files of TCGA, we used the data from GDC. GDC provided the TCGA data in Bam format, *raw counts*, FPKM, and FPKM-UQ files. Since we knew the correlation between the entropy and the 5-years *overall survival* (OS) for nine cancer types as established from TCGA RSEM-UQ (Conforte et al., 2019), the validation challenge was (i) to normalize the GDC *raw counts* files (we characterized this step as $RPKM_{upper}$, see the description below) from tumors of the nine cancer types; (ii) to compare the $RPKM_{upper}$ normalization to the TCGA RSEM-UQ for critical value, number of up-regulated genes, and the correlation between entropy and 5-years OS as well as targets; (iii) to compare $RPKM_{upper}$, TCGA RSEM-UQ and GDC FPKM-UQ for critical value, number of up-regulated genes, the correlation between entropy and 5-years OS, and targets, and (iv) to optimize $RPKM_{upper}$ by log transformation for target selection given the maximization of the correlation coefficient of the relationship between entropy and 5-years OS. Having this process validated, it might be applied to any method of read counting from *fastq* file by read mapping. This process is summarized in **Figure 1**.

As TCGA, GDC uses the RSEM methodology to map reads to reference genes. Here, instead of using the human genome sequence GRCh38.d1.vd1[14], we used the proteins sequences from UniprotKB as a reference. Since only about 80% of the proteins from the EBI's interactome referenced by UniprotKB matched the *consensus coding sequences* (CCDS)[15] of Ensembl, we decided to map reads in *fastq* files directly with the proteins sequences of the intact-micluster interactome using BLASTx. Thus, in the first instance, the exercise of validation concerned the processing of *raw counts* into RPKM-UQ output.

For *raw count* normalization, we used a modified version of the RPKM formula (5):

$$RPKM = \frac{RC_g * 10^9}{RC_{pc} * L} \tag{5}$$

where:

RC_g: Number of reads mapped to the gene;

RC_{pc}: Number of reads mapped to all protein-coding genes;

L: Length of the coding sequence in base pairs.

RPKM is relative to the total number of reads, which is a linear expectation. Quantile normalization (Bolstad et al., 2003) forces the distribution of the normalized data to be the same for each sample by replacing each quantile with the average quantile across all samples. Instead, one may focus on a specific quantile. For instance, the upper quartile normalization (Bullard et al., 2010) divides each read count by the 75th percentile of the read counts in its sample. However, the gene frequency (y) according to the gene expression (x) follows a power law (the relationship of log(y) and log(x) is linear, data not shown) (see also Balwierz et al., 2009; Awazu et al., 2018). RPKM, as defined in formula 5, does not take the non-linearity associated to large expression level into account. By contrast, the *upper quartile* normalization enables us to take the non-linearity associated with extreme expression values into account. Formula 5 can be written as formula 6:

$$RPKM_{upper} = \frac{RC_g * 10^9}{L * (RC_{pc} - (\delta * RC_{pc}))} \tag{6}$$

where δ is a tuning factor.

For δ = 0, formula 6 is equivalent to RPKM (formula 5) and for δ = 0.25, it is equivalent to a *upper quartile* normalization. In this work, we used δ = 0.05 because it optimized the coefficient of correlation between entropy and 5-years OS.

It appeared that in addition to the TCGA RSEM-UQ (accessed in 2016), GDC (accessed in March 2020) implemented a correction for false positive minimization (Anders and Huber, 2010; Love et al., 2014; Holmes and Huber, 2019). The result of this minimization is a flatten power law of gene expression with an effect similar to that of formula (7):

$$LogNorm = C * x_i * (log_b \left(log_b(x_i + 1)\right) + 1) \tag{7}$$

where:

C: is a constant that was set to 20 to optimize the coefficient of correlation of the relationship between entropy and 5-years OS;

x_i: is the $RPKM_{upper}$ value of the i_{th} element;

b: is the base of the logarithm, which was set to 1.1.

As can be seen from formula 7, the FPKM-UQ output follows a *log-log* relationship except for the variance that is stabilized by a Bayesian process.

For assessing the efficiency of TCGA *raw counts* processing according to formula 6, we tabulated the sample size of subnetworks of up-regulated genes as well as the critical values obtained for PDF = 0.975. This process was performed by calculating $RPKM_{upper}$ on the *raw counts* available from GDC, and compared the critical values to those obtained from GDC FPKM and TCGA RSEM-UQ. We also compared the correlation between entropy and 5-years OS obtained with *raw counts* normalized with $RPKM_{upper}$ to that obtained by using the TCGA RSEM-UQ. Finally, we compared the most relevant targets obtained from both processes.

In the case of the GDC FPKM-UQ, one more step was necessary since the *raw counts* sequentially processed through formula 6 and 7 had to be compared to FPKM-UQ data available from the GDC portal. Again, we compared the performance of processing *raw counts* with formula 6 and 7 to GDC FPKM-UQ data considering (i) the critical values for PDF = 0.975, (ii) the subnetwork size of up-regulated genes, (iii) the correlation of entropy vs. 5-years OS, and (iv) the list of most relevant targets obtained through both processes.

Finally, we also compared the performance of sequentially processing *raw counts* through formula 6 and 8 (formula 8 is derived from Cloonan et al., 2008) by using the same measures as just described (i to iv). We applied this formula because we

[14]https://gdc.cancer.gov/about-data/data-harmonization-and-generation/gdc-reference-files

[15]https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi

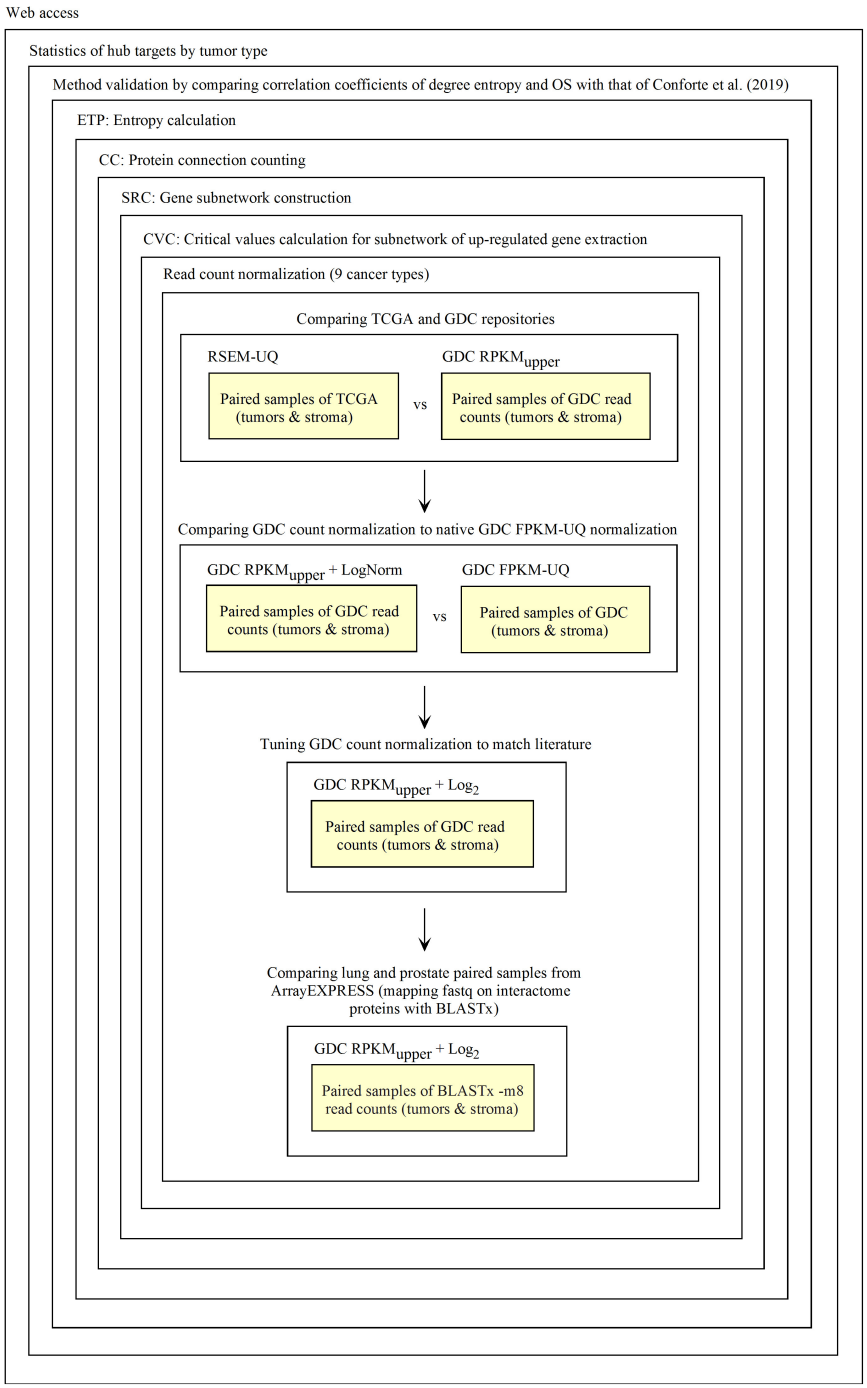

FIGURE 1 | Process of Galaxy workflow validation.

noticed that it optimized the coefficient of correlation of the relationship between entropy and 5-years OS.

$$Log2 = x_i \left(log_b(x_i + 1) \right) \tag{8}$$

where:

x_i: is the $RPKM_{upper}$ value of the i_{th} element;

b: is the base of the logarithm, which was set to 2.

Galaxy Scripts

Galaxy is a scientific open-source workflow platform that aims at helping users to perform repetitive and complex operations over large datasets. With Galaxy, users can visually create processing pipelines reproducing the data flow over programs and datasets that are viewed as interconnected box objects. Additionally, Galaxy is written in Python and JavaScript, but has an XML like

interface able to transfer the processing flux to other languages. Galaxy comes with a rather large initial set of tools that can be added to the desktop according to simulation demands. Internally, every Galaxy tool is made up of a XML file that describes its functionalities and interface. Once XML interfaces are programmed, Galaxy is very simple to operate in an object-oriented mode by linking input data with scripts together.

By means of a specific script (see below), Galaxy can store data in MongoDB, which is a non-relational object-oriented database (NoSQL) (Bradshaw et al., 2019). MongoDB can be accessed through Angular, which serves as a frontend framework for users (the physician or/and technician operating the system) (Fain and Moiseev, 2018).

As outlined in the introduction of this report, our Galaxy workflows are derived from the agglomeration of Perl scripts (except for CVC.py) that were written for previous reports (Carels et al., 2015a; Conforte et al., 2019). These tools are as follow:

(1) *Count Connections* (CC) counts the number of connections that each protein has with their neighbors in a subnetwork of up-regulated genes. CC is an intermediate step to compute the entropy.
(2) *Critical Value Calculation* (CVC) computes a critical value according to the normal distribution that fits the observed data and a probability level informed by the user. All genes with expression values above the critical value, used here as a threshold, are considered as up-regulated.
(3) *Differentially Expressed Genes List* (DEGL) computes de differential gene expression between RNA-seq data from tumoral and control samples (tumor minus control).
(4) *Entropy Calculation* (ETP) computes the Shannon entropy corresponding to a subnetwork. Here, we typically considered the subnetworks of genes that are up-regulated in tumors.
(5) *Translation of Gene Symbol into UniProt KB accession numbers* (GS2UP). Former TCGA data files identified genes by gene Symbol, while the interactome from EBI (the intact-micluster.txt file) uses UniProtKB accession numbers. GS2UP translates the gene symbols to UniProtKB accession numbers to build the subnetwork of up-regulated genes.
(6) *Translation from Ensembl into UniProt KB accession numbers* (Ensembl2UP). GDC data files identify genes by reference to Ensembl, while the interactome from EBI (the intact-micluster.txt file) uses UniProtKB accession numbers. Ensembl2UP translates the Ensembl to UniProtKB accession number to build the subnetwork of up-regulated genes.
(7) *Protein To Total Connections Sorted* (PTTCS) sorts the file of malignant up-regulated genes according to the level of connectivity found for their respective protein in descending order.
(8) Subnetwork Construction (SRC) computes a subnetwork of proteins based on a gene list by reference to the intaractome; here, the gene list is typically the list of up-regulated genes.
(9) *Reads Per Kilobase Million – Upper Normalization* ($RPKM_{upper}$) computes de normalization of RNA-seq data according to formula 6.
(10) *Double Logarithm Transformation* (LogNorm) computes de normalization of $RPKM_{upper}$ data according to formula 7.
(11) *Base 2 Logarithm Transformation* (Log2) computes de normalization of $RPKM_{upper}$ data according to formula 8.
(12) *PTTCS to MongoDB* (P2M) computes the data storage within MongoDB.

These tools can be downloaded from GitHub: https://github.com/BiologicalSystemModeling/Theranostics under the MIT License, however, the concept of theranostics based on this approach is under the regulation of intellectual property number BR1020150308191 for Brazil.

Pipeline Scaling

To investigate how the pipeline scales, we processed the GDC raw counts data using an AMD Ryzen 9 3900X (4.6 GHz) CPU with 20 threads dedicated to Galaxy and 64 GB RAM. First, we chose LUSC and PRAD tumors as representing high entropy (low OS) and low entropy (high OS) cancer types, respectively. In these two cases, we could exactly compare their scaling until 45 patients by increments of five. For STAD, LIHC, THCA, and KIRC, we measured the processing time for only two patient numbers (15 and 25). We also analyzed the statistical significance of the difference in processing speed observed for entropy and PTTCS pipeline for 25 patients with the Student's *t*-test. Considering the pipeline for hub diagnosis from BLASTx output, we only had access to a small number of patients, which limited the power of the experiment. We compared 3, 6, 9, 12 patients in PRAD and NSCLC from ArrayEXPRESS (see above).

Web Application

As outlined in the introduction, we aimed at releasing a tool based on a phenotyping approach for the rational therapy of cancer. At the moment, the current approach of cancer therapy is still largely based on mutation mapping (genotyping approach), but the potential benefits of integrating RNA-seq data must be considered and this is the purpose of this report.

When producing a bioinformatic application, it is necessary to validate it according to some objective criterion. As presented in the previous section, we chose degree entropy as such a criterion for the validation of the Galaxy pipeline. Galaxy enabled us to test the performance of several configurations for optimizing the correlation between the degree entropy of up-regulated subnetworks and the patient's 5-years overall survival.

However, a website is necessary to make this tool available to the medical community and its development makes part of another step of validation that is its acceptance by professionals. Below, we briefly describe the technologies that we used to build the web site and then described how we implemented them through forms for data submission.

MEAN Stack

Both MongoDB and Angular are part of the MEAN stack (MEAN for M of MongoDB, E of Express.js, A of Angular, and N of Node.js). The use of MongoDB with Node.js, its native driver, is facilitated by the Mongoose[16] library. Mongoose, amongst other benefits, allows (i) the use of JavaScript as a programming language, which save the need for database programming, (ii) the modeling of data before their saving into MongoDB, and (iii) the *horizontal scaling*[17], which means that one can expand storage capacity without the need of multiple structural changes. This last feature decreases the cost of prototyping and expansion. It also enables one to work with several database connections simultaneously.

Node.js is part of the MEAN stack that we used to build the backend of the web application; it is the server used to connect the database and the frontend. Essentially, Node.js is a framework that is used to create servers and has its own HTTP handler (Holmes and Herber, 2019), which eliminates the need of other intermediate libraries.

The MEAN also included Express.js, a JavaScript-based library whose purpose is to facilitate the exploration of the Node.js functionalities (e.g., creating routes).

In addition to JavaScript, Angular also allows programming in TypeScript, which includes the concept of *variable type* and a set of internal libraries (e.g., RxJS for asynchronous programming). Furthermore, Angular offers compatibility with many web development libraries, such as Bootstrap, jQuery, and Forms.

MEAN stack elements have JavaScript as a common programming language and *JavaScript Object Notation* (JSON) as a common file exchange format. Except for Angular which is a frontend technology, MongoDB, Express.js, and Node.js run on the server-side, as so they are generally classified as the 'backend' of a web application (Holmes and Herber, 2019).

Our web application has been deployed in a cloud environment using Heroku[18,19] by implementing the MEAN stack (Holmes and Herber, 2019). The version of Angular that we used here was CLI 8.3.23. In addition to those technologies, we were also using NPM libraries designed to support the MEAN stack. We used JavaScript for interfacing with MongoDB, Express.js, and Node.js as well as several free packages available in NPM to support these technologies[20]. For instance, we used *Visual Studio Code* (version 1.48) as a programming platform and *Avast Secure Browser* as a testing browser. Avast provides a built-in test system for small devices such as smartphones.

Angular

After compilation, Angular generates *Single Page Applications* (SPAs), which means that the code is sent to the browser at once when the user accesses the page for the first time. The main benefit of this approach is to create *dynamic pages*, improving the navigation experience to the frontend user. Angular speeds up the server–client communication by avoiding multiple client accesses and enabling complex calculations as well as data validations within the client browser. Moreover, the main difference of SPAs compared to a classic web application based on PHP (i.e., *static pages*) is that it does not load the page when one changes from page to page since all the code is already on the browser. Therefore, the main benefits of Angular are that (i) heavy calculations can be performed on the frontend side, which can alleviate the computing charge on the server; (ii) pre-validated data may be submitted to the server, avoiding the need for *back and forth* validation process; (iii) TypeScript (a superset of JavaScript) has the structures of a conventional programming language with powerful build-in libraries (e.g., RxJS), which enables the performance of scientific calculations on the frontend side if needed.

We also took advantage from the Angular library called *Angular Material*[21], which allows predefined functions such as forms and themes. Angular Material can be used either within the HTML language as predefined tags or within TypeScript for dynamic pages (e.g., for Reactive forms). We used Angular Material within TypeScript since it provides much more programming freedom, e.g., form validation.

Node.js

One of the key features of Node.js is that it allows the usage of JavaScript (or TypeScript) on the server-side. Until then, JavaScript was restricted to browsers and this progress has been possible due to the V8 Engine that compiles JavaScript code to native machine code at runtime. We used the NPM repository to install and manage all the Node.js (version 10.16.3) packages.

Node.js applications are *stateless*, which means that they do not keep information about the user stored locally and for that reason only require low amount of local RAM. Node.js applications are also single thread, which means that they do not stop the main thread as they result from users' interactions.

We chose the *JSON Web Token* (JWT) approach to save the user information temporally on the frontend. JWT is an encoded string used when the frontend communicates with the server. The benefits of JWT are (i) that it carries a server signature, which must match whenever the user tries to communicate with the server, and (ii) that an expiration date may be set, which implies token refreshing.

Express.js

Express.js is a library whose purpose is to facilitate the exploration of the Node.js functionalities (e.g., creating routes and servers). Here, we used Passport.js[22] together with Express.js (version 4.16.1) to build user sections as described by Holmes and Herber (2019).

MongoDB

MongoDB can be accessed through Angular using Node.js as server; Angular serves as a frontend framework for users (Fain and Moiseev, 2018). MongoDB is *horizontally*

[16] https://mongoosejs.com/docs/

[17] https://docs.mongodb.com/manual/sharding/

[18] https://www.heroku.com/

[19] http://teranostico.herokuapp.com/

[20] https://www.npmjs.com/package/repository

[21] https://material.angular.io/. Accessed on 14/10/2020.

[22] http://www.passportjs.org/. Accessed on 14/10/2020.

expandable[23], which enables to expand storage capability without extensive physical changes. This feature decreases the cost of prototyping and posterior expansion. Another interesting property of MongoDB is the *MongoDB Atlas*[24], which provides cloud storage.

The usage of MongoDB with Node.js is facilitated by the Mongoose[25] library. Mongoose, amongst other benefits, allows (i) the usage of JavaScript as a programming language, which saves the need for database programming, (ii) the modeling of data before their storage into MongoDB, and (iii) the easier exploration of the MongoDB horizontal scaling capability[26].

Angular Flex-Layout

According to Fain and Moiseev (2018), we used a single code to implement *Responsive Web Design* (RWD) to optimize maintenance costs. This strategy allows the user interface layout to change in response to the device screen size (desktop or cell phone). RWD allows the interface simplification on small devices by limiting the display of extra-small devices to key functions (see **Supplementary Figure 1** for screen size and Angular screen size settings).

We tested the responsiveness of our portal on a desktop computer using the built-in developer tool of Avast Secure Browser. We also tested it on the following devices: Moto G4, Galaxy S5, Pixel 2, Pixel 2 XL, iPhone 5/SE, iPhone 6/7/8, iPhone 6/7/8 Plus, iPhone X, iPad, iPad Pro. However, the Avast Secure Browser simulator does not necessarily consider the operating system, and it may give an unexpected display in uncommon devices.

Passport.js

For creating the user section, we used Passport.js[27]. Its main benefits are the possibility of (i) creating customized login system or use pre-defined ones, such as those of Facebook, for example; and (ii) using it with JWT tokens due to their built-in libraries that facilitate their use. To implement JWT within Passport.js, we used *express-jwt*[28], which allows the validation of JWT tokens, including expiration date and abnormal tokens.

Forms

The function of the patient main form is to collect and to store basic information regarding the patient and its tissue samples for genetic analysis. This information is necessary for the posterior retrieval from the system database of patients' medical records. Patient data are central to the system since they articulate genetic analyses with medical records that must be encrypted (e.g., patient name, mother's name, and patient id). The patient data collected through the main form of the frontend are stored together with genetic data from the backend within MongoDB.

The request for a genetic exam is of key importance when it comes to the service provided. When physicians send tumor samples, they will be asked to request their gene expression analysis and provide patient information as well as medical records (see **Supplementary Figure 2**).

The outcome form has such as (i) details of the treatment applied, (ii) treatment benefits, (iii) whether the gene expression-based recommendations were followed, and so forth (see **Supplementary Figure 3**). The outcome form is essential for establishing case statistics.

Angular provides two options when it comes to forms: *Template-Driven Forms* and *Reactive Forms*[29]; we used the latter. The main reason for this choice was that this option provides (i) a set of built-in routines for form validation, including error messages that can easily be shown on the frontend, and (ii) the possibility of building its own customized error handling routines. By error, we mean any input to the form fields that does not fit what is expected, e.g., e-mail out of the format or password that does not match. We were also using form validators that communicate with the server on the background side to check data consistency.

Additionally, we used FormBuilder[30] that is an Angular service used for the programming of Reactive form. With FormBuilder, one can construct JSON objects (our data format), validate the inputs of the forms individually or as a group, and other functionalities.

Encryption, Decryption, Hashing, and JWT Coding

Since we are dealing with potentially sensitive information, we followed standard practices to protect the information submitted to the system and stored on our database. In the current stage of development, we are using standard libraries, which can be replaced by more secure ones as soon as the platform scale up. In the current version, we are using three different approaches to protect information from potential unauthorized accesses: (i) encryption/decryption, (ii) hashing, and (iii) JWT (e.g., communication with API[31]). For encryption/decryption, we are using the library *CryptoJS*.[32] The 'secret' is kept on the server using a library known as *dotenv*[33], which is largely used to store sensitive information in Node.js applications. For hashing, we are using the library *bcrypt*[34] in the following configurations: *bcrypt.genSalt*(*10, callback*), the first argument is the size of the *salt* and the second is the function for hashing.

[23]https://docs.mongodb.com/manual/sharding/. Accessed on 14/10/2020.

[24]https://www.mongodb.com/cloud/atlas. Accessed on 14/10/2020.

[25]https://mongoosejs.com/docs/. Accessed on 14/10/2020

[26]https://docs.mongodb.com/manual/sharding/. Accessed on 14/10/2020.

[27]http://www.passportjs.org/. Accessed on 16/10/20.

[28]https://www.npmjs.com/package/express-jwt. Accessed on 16/10/20.

[29]Components in the Angular realm is a set of three files: CSS (appearance-related), TS (typescript, coding), and HTML (classical static page design file). A page is built from at least one component, which can independently interact with each one of the others (see Fain and Moiseev, 2018 for a more detailed discussion).

[30]https://angular.io/guide/reactive-forms

[31]Application Programming Interface. These routines are designed to access the database following some pre-defined rules such as token authentication.

[32]https://www.npmjs.com/package/crypto-js

[33]https://www.npmjs.com/package/dotenv

[34]https://www.npmjs.com/package/bcrypt

The code for the web site can be downloaded from GitHub: https://github.com/Teranostico under the MIT License.

RESULTS

Galaxy Pipeline

We validated and automated the process published by Conforte et al. (2019). Thus, one sought to reproduce the results obtained by Conforte et al. (2019) when the pipeline was fed with the same data (TCGA RSEM-UQ). We indeed succeeded to reproduce the correlation $r = -0.68$ between entropy and patient's 5-years OS for a probability of $p = 0.975$ in the determination of up-regulated genes, which allowed us to test whether the maximization of r really occurred for $p = 0.975$. To meet this challenge, we measured the correlation coefficient for $p = 0.97$ and $p = 0.98$, and found $r = -0.53$ and $r = -0.60$, respectively. The automated workflow is given in **Figure 2A**.

As shown in **Figure 2A**, the *input data collection* represents a collection of paired samples (tumors identified as 01A and control identified as 11A) with the same list of genes (identified by gene symbol) for each patient of the TCGA database. Following the processing flux, the gene symbols are transformed into UniprotKB accession numbers (GS2UP) to perform the subtraction of the control RNA-seq expression data from that of the tumor (DEGL). The calculation of the critical value that identifies up-regulated genes is performed by the Python script CVC. The critical value is calculated according to a probability level chosen by the user and is used by the script SRC for extracting the list of up-regulated genes. This list is used by the CC script for counting the connections at each vertex of the subnetwork of up-regulated genes. The connection count at each vertex is necessary for computing the Shannon entropy of the tumor subnetwork of up-regulated genes by the ETP script.

We validated the pipeline with the GDC *raw counts* comparing their $RPKM_{upper}$ to the TCGA RSEM-UQ (**Figure 2B** without the log transformation step). First, we computed the *raw counts* according to $RPKM_{upper}$ excluding BRCA and LUAD because of inconsistencies between file names available for FPKM-UQ and *raw counts*. In both BRCA and LUAD, cleaning samples for perfectly matched files led to sample size below $n = 20$, which may bias comparison (sample size is considered to be statistically trustworthy from at least $n = 30$ and needs correction below this threshold). When we compared the critical values for $p = 0.975$ considering the *raw counts* normalized with $RPKM_{upper}$ (**Table 2**, column GDC $RPKM_{upper}$), we found values similar to those obtained by processing TCGA RSEM-UQ data (**Table 2**, column TCGA RSEM-UQ).

We found that critical values for $p = 0.975$ of GDC FPKM-UQ were ∼5 times larger (**Table 2**, column GDC FPKM-UQ), on the average (**Figure 2B** without normalization and log transformation steps), than those of TCGA RSEM-UQ (**Table 2**, column TCGA RSEM-UQ and GDC $RPKM_{upper}$). This difference is due to the processing update performed during the data transfer from TCGA to GDC portal involving the flattening of the differential gene expression distribution.

When we successively computed GDC *raw counts* with formula 6 ($RPKM_{upper}$) and 7 (LogNorm), we found critical values for $p = 0.975$ (**Table 2**, column GDC $RPKM_{upper}$ + LogNorm) close to that of GDC FPKM-UQ (**Table 2**, column GDC FPKM-UQ), suggesting a similar behavior of differential gene expression flattening as the one applied by the GDC data processing (**Figure 2B**).

The comparison of the size of subnetworks of up-regulated genes in tumors is given in **Table 3**. The difference of subnetwork size between GDC FPKM-UQ and GDC $RPKM_{upper}$ + LogNorm samples, on one hand, and TCGA RSEM-UQ and GDC $RPKM_{upper}$ samples, on the other hand, raised the question of whether the large subnetwork size of GDC FPKM-UQ and GDC $RPKM_{upper}$ + LogNorm might be trusted.

The subnetwork sizes obtained by successively processing GDC *raw counts* with formula 6 and 8 (**Table 4**, column Node number) were smaller and more realistic, representing between ∼2% and ∼5% of the human proteome.

As explained above, we did not consider BRCA and LUAD for comparison between $RPKM_{upper}$ and FPKM-UQ. However, the FPKM-UQ correlation plot was similar to that of other authors (data not shown).

The features of the linear regression between the subnetwork entropies and the 5-years OS are given in **Table 5** for the different pipeline configurations tested here.

Interestingly all the combinations involving $RPKM_{upper}$ of **Table 5** resulted in a larger slope of the regression line; in other word, they resulted in an increased statistical significance of the regression line.

Compared to GDC $RPKM_{upper}$ (**Figure 3A**), the introduction of the LogNorm in the workflow of **Figure 2B** resulted in a systematic shift of entropies by as much as ∼1.5 bit toward larger values (in the range of 3.6–4.0 compared to 2.0–2.5 in Conforte et al., 2019), which denote a larger subnetwork of up-regulated genes with larger number of hubs as a consequence of the distribution flattening of differential gene expression. The correlation obtained by successively processing *raw counts* with $RPKM_{upper}$ and LogNorm (**Figure 3B**) was similar ($r = -0.86$ without BRCA and LUAD) to that obtained with GDC FPKM-UQ ($r = -0.76$ without BRCA and LUAD) (**Figure 3D**). Finally, it is the correlation obtained by successively processing *raw counts* with formula 6 and 8 (**Figure 3C**) that showed the best correlation coefficient and slope of the regression line ($r = -0.91$).

The effect of LogNorm on distribution flattening of differential gene expression when comparing $RPKM_{upper}$ to $RPKM_{upper}$ + LogNorm was similar to that observed when comparing TCGA RSEM-UQ (**Figure 4A**) to GDC FPKM-UQ (**Figure 4D**), respectively.

When we compared the correlation coefficient according to p for GDC FPKM-UQ data, we obtained $r = -0.758$, $r = 0.763$, and $r = 0.477$ for $p = 0.95$, $p = 0.98$, and $p = 0.99$, respectively. This result shows that the maximum of r was associated with $p = 0.98$, but the difference with $p = 0.975$ was only 0.002 units of the correlation coefficient, which confirmed that the peak around the maximum of r was less sharp for GDC FPKM-UQ than for TCGA RSEM-UQ since it spreads over $p = 0.95$ and $p = 0.98$.

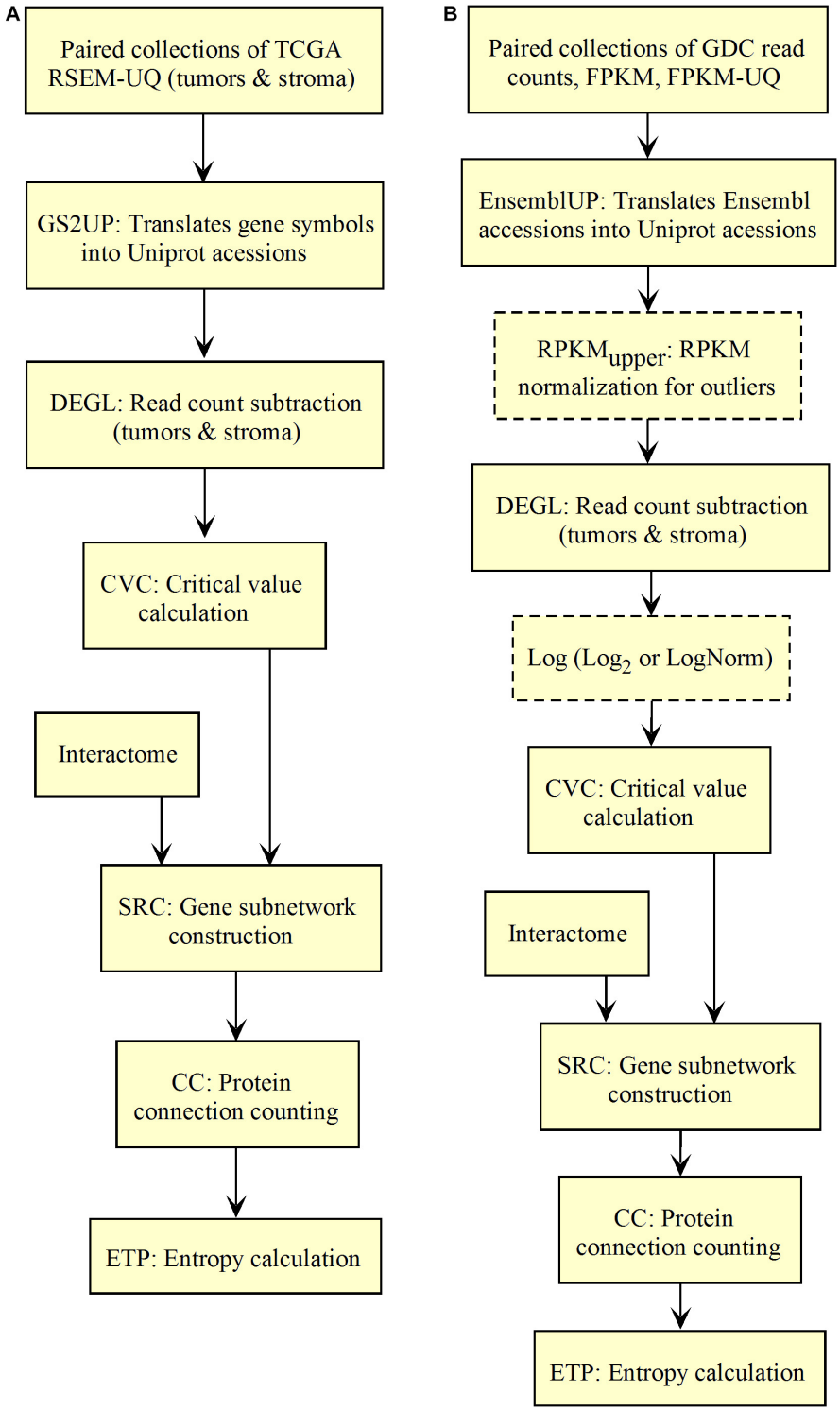

FIGURE 2 | Workflow for the validation of the correlation between the entropy of the subnetworks of up-regulated genes from different tumors and their respective 5-years OS. **(A)** TCGA. **(B)** GDC.

The flattening of the correlation peak according to the probability density appeared as a consequence of the probability density distribution shape. The distribution of FPKM-UQ values was flatter in GDC FPKM-UQ (**Figures 4D–F**) compared to TCGA RSEM-UQ (**Figures 4A–C**), which is reflected by larger critical values associated with GDC (**Table 2**). The validation of

TABLE 2 | Critical values of probability density for $p = 0.975$.

Cancer	GDC $RPKM_{upper}$		TCGA RSEM-UQ		GDC $RPKM_{upper}$ + LogNorm		GDC FPKM-UQ	
Type	Av.	StDev	Av.	StDev	Av.	StDev	Av.	StDev
PRAD	2661.73	498.89	2566.88	507.38	15558.79	1053.56	15809.77	779.91
LUAD	2897.95	437.50	3138.07	313.74	15720.85	1221.94	16340.59	860.48
LUSC	3532.06	426.30	3527.89	429.98	15775.55	857.31	16161.27	730.71
BRCA	3211.72	434.50	3024.87	465.83	15346.96	664.95	15923.64	682.80
KIRC	3133.16	236.39	3162.20	363.44	15820.34	742.76	16310.36	604.57
KIRP	3084.69	365.17	3089.64	390.28	15482.35	905.77	16165.59	597.30
THCA	2610.75	313.49	2590.59	406.12	14876.38	1089.35	15559.35	713.54
STAD	3330.13	444.58	3273.89	470.64	16511.00	865.11	16473.27	718.44
LIHC	3085.76	474.40	3409.36	468.48	16235.74	1087.23	15639.15	801.43
Average	3060.88	403.47	3139.90	420.79	15703.11	943.11	16042.55	721.02
St. Dev.	298.36	83.63	299.07	56.29	479.01	181.82	324.01	86.51

TABLE 3 | Size of subnetwork (vertex number) of genes up-regulated in tumors for a probability density of $p = 0.975$.

Cancer	GDC $RPKM_{upper}$		TCGA RSEM-UQ		GDC $RPKM_{upper}$ + LogNorm		GDC FPKM-UQ	
Type	Av.	StDv.	Av.	StDv	Av.	StDv.	Av.	StDv.
PRAD	269.19	62.01	254.20	40.66	5046.23	1209.45	4029.75	499.89
LUAD	290.35	58.66	276.35	49.07	4973.16	1203.25	4779.27	401.19
LUSC	345.21	48.50	317.12	48.33	5824.63	904.38	4981.60	460.49
BRCA	311.55	46.50	286.50	42.16	5305.85	1219.24	4816.61	361.22
KIRC	332.28	42.37	328.10	52.85	5117.83	881.77	4556.75	294.80
KIRP	313.48	49.64	303.22	41.14	4983.68	1136.93	4678.77	305.26
THCA	256.52	47.31	276.95	57.44	4016.13	948.02	4142.73	387.09
STAD	341.67	52.90	276.59	51.66	6773.41	928.62	4764.48	351.76
LIHC	352.74	68.99	256.24	85.31	7007.28	143.05	4522.08	400.83
Average	312.55	52.99	286.14	52.07	5449.80	1096.08	4585.78	384.72
St. Dev.	34.34	8.57	25.49	13.72	942.86	189.53	315.90	66.69

the mapping process of reads on the EBI interactome proteins needed similarity comparison of *fastq* files using BLASTx. We performed this validation by recycling the components of **Figure 2** for processing RNA-seq data as shown in **Figure 5**.

The workflow shown in **Figure 5** needed to be fed with BLASTx outputs. After mapping reads to their respective protein sequences in the interactome, both tumor and control *raw count* files were normalized ($UTCENG_{upper}$) according to their coding sequence size ($RPKM_{upper}$ step) and expression level using formula 2. The rest of the pipeline is as in **Figure 2** except for the last step of sorting by decreasing level of connection (PTTCS) and data storage in MongoDB (P2M).

The list of top-n connected up-regulated hubs is released as output data from the workflow, and stored in MongoDB (**Figure 5**) together with the patient's clinical data. These data can be formatted as a medical report by the JavaScript code within the web page according to the user request.

Considering the entropies of subnetworks of NSCLC up-regulated genes (x_1 = PRJNA320473) and PRAD (x_2 = PRJEB2449), the u_{obs} calculated with formula 3 with $\bar{x}_1 = 2.99475$ and $\bar{x}_2 = 1.66472$, respectively, as well as $SCE_1 = 10.31347$ and $SCE_2 = 6.70566$, respectively, was 5.00748. Since k was found to be 29.06411 ($\sim$29) for the sample sizes considered, the theoretical values of t for $p = 0.975$ and $p = 0.999$ were 2.045 and 3.396, respectively. Because $u_{obs} > t_{th}$, we rejected the null hypothesis of average equality for NSCLC and PRAD and concluded that the entropy of NSCLC was

TABLE 4 | Critical values of $RPKM_{upper}$ + Log2 for a probability density of $p = 0.975$ and vertex number of subnetworks of genes up-regulated in tumors.

Cancer	Critical value		Vertex number	
Type	Average	StDev	Average	StDev
PRAD	7359.58	1019.74	884.69	248.58
LUAD	7985.00	1105.05	946.58	219.40
LUSC	9325.45	1187.62	1244,35	232,65
BRCA	8398.81	1232.49	1087.20	240.74
KIRC	8335.51	561.98	1035.82	143.58
KIRP	8299.87	777.86	1014.35	206.89
THCA	7210.95	706.19	775.96	140.05
STAD	9173.28	1154.74	1264.93	249.58
LIHC	8398.81	1232.49	1235.50	294.36
Average	8276.36	997.58	1054.38	219.53
St. Dev.	707.07	251.53	171.09	50.27

TABLE 5 | The features of the linear regression between the entropy and the 5-years OS for $p = 0.975$.

Normalization method	Coef. Correl. (with BRCA + LUAD)	Coef. Correl. (without BRCA + LUAD)	Regression (without BRCA + LUAD)
GDC RPKM	–0.36	–0.55	–
GDC $RPKM_{upper}$	–0.68	–0.86 (**Figure 3A**)	$y = -0.0084x + 2.717$
GDC $RPKM_{upper}$ + LogNorm	–0.67	–0.85 (**Figure 3B**)	$y = -0.0090x + 4.473$
GDC $RPKM_{upper}$ + Log2	–0.69	–0.91 (**Figure 3C**)	$y = -0.0096x + 3.460$
GDC FPKM	–0.11	–0.13	–
TCGA FPKM-UQ*	–0.68	–0.64	$y = -0.004x + 2.507$
GDC FPKM-UQ	–0.71	–0.76 (**Figure 3D**)	$y = -0.0039x + 4.025$

*See *Conforte et al., 2019*.

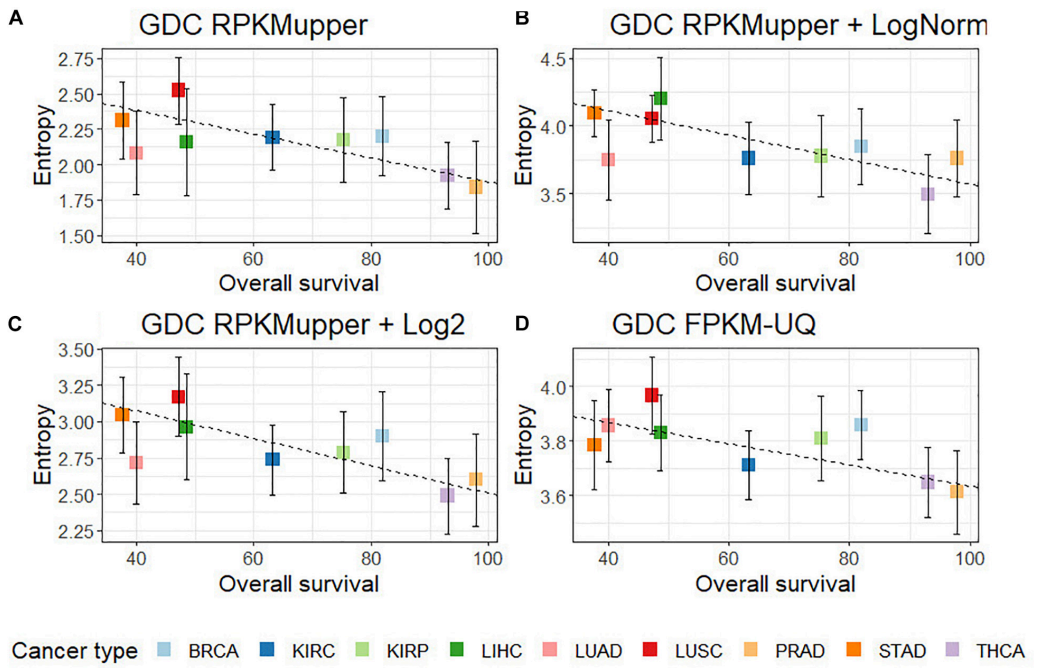

FIGURE 3 | Correlation of subnetwork entropies vs. 5-years OS for $p = 0.975$. **(A)** GDC $RPKM_{upper}$. **(B)** GDC $RPKM_{upper}$ + LogNorm. **(C)** GDC $RPKM_{upper}$ + Log2. **(D)** GDC FPKM-UQ.

significantly larger than that of PRAD. This result is in agreement with the negative correlations of **Figure 3** and validates the pipeline here presented.

As the methodology was validated, it could be used for the diagnosis of the top-n most connected proteins within the list of up-regulated genes in the tumor compared to the stroma. It is important to underline that the entropy was used only for the purpose of methodology validation.

A pipeline to identify the connection hubs is given in **Figures 6A,B**, where the purpose of PTTCS is to compare up-regulated genes to the list of vertex connections in the interactome to rank them in decreasing order of connection number in the output file. *A priori*, top-20 most connected proteins among the up-regulated genes of tumors should be enough to design a personalized treatment. However, this number depends on drug availability.

The comparison of the most relevant targets associated with the different normalization methods applied in this report is shown in **Table 6**. **Table 6** reports the number of tissues (# column) where the gene of a given protein (Acc column) was up-regulated among nine different tumors. For illustration, we only kept genes up-regulated in at least 70% of tumor samples of each cancer type (pink). The colors in the first column report for the targets that are common between different sections (A to E) of **Table 6** (turquoise is for the genes common to **Tables 6A–E**; blue is for the genes common to **Tables 6A,B,D,E**; yellow is for the genes common to **Tables 6A–C,E**; mallow is for the genes common to **Tables 6A,B,E**; and green is for the genes common to **Table 6D,E**).

Tables 6A–E show that the most relevant targets are largely shared among methods. In **Tables 6C,D**, target personalization according to the tumor was lower than in **Tables 6A,B,E**. Because of the larger average network size that it produced, the normalization with $RPKM_{upper}$ + Log2 (**Table 6E**) showed a larger targets number than TCGA RSEM-UQ and GDC $RPKM_{upper}$ (**Tables 6A,B**), similar to those of **Tables 6C,D**

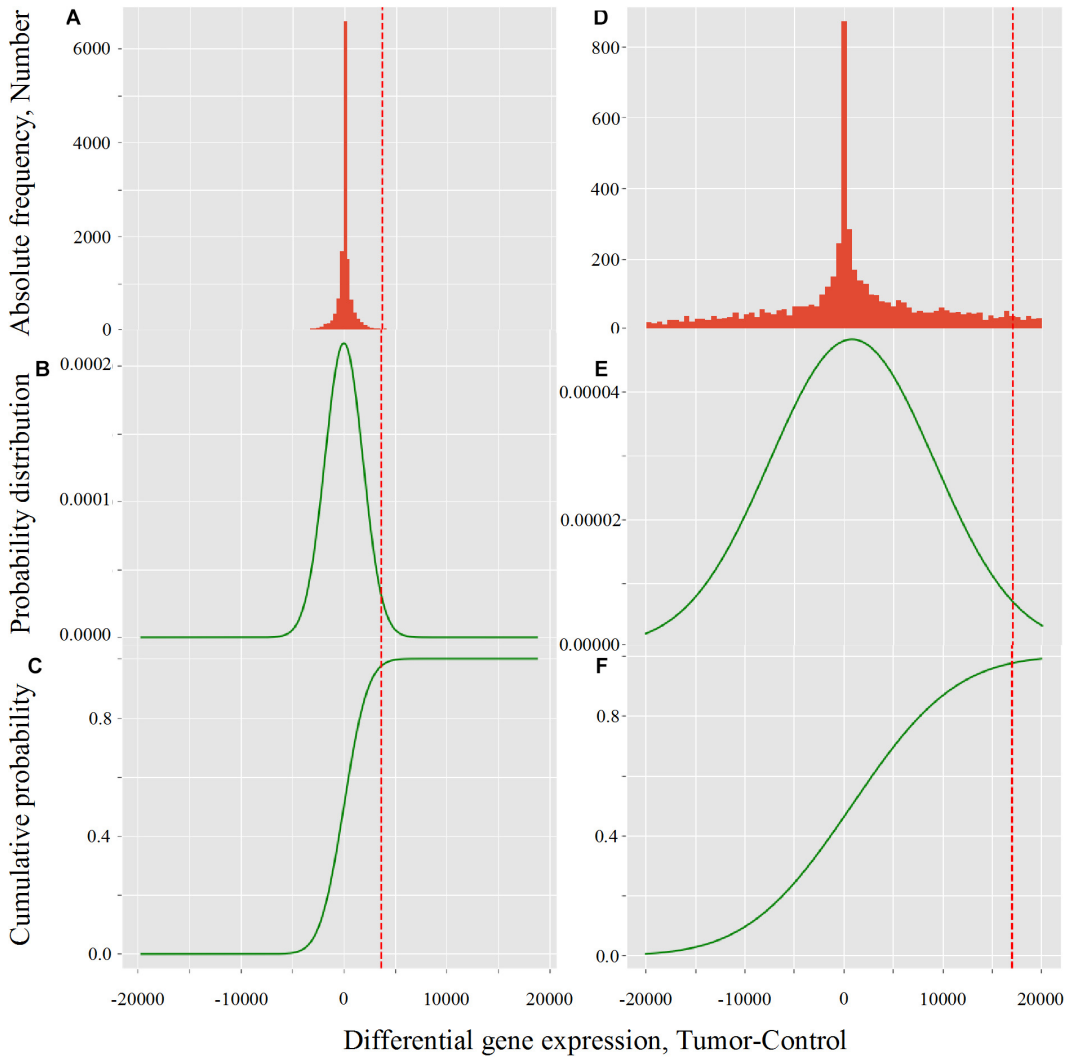

FIGURE 4 | Critical value calculation (red dot line) by CVC script in TCGA **(A–C)** and GDC **(D–F)** in LUSC (TCGA-22-4593 sample). **(A,D)** Histogram of observed differential gene expression distribution (tumor-control) of genes. **(B,E)** Function of density of probability. **(C,F)** Function of cumulated probability. The critical values were 3,633.8 and 17,042.9 for TCGA and GDC, respectively.

but with a larger level of tumor personalization. Because of the reasonable size of subnetworks and the best correlation relationship between entropy and the 5-years OS it produced, the successive processing through $RPKM_{upper}$ and Log2 normalization was considered here as the best compromise.

Scaling Analysis

The analysis of LUSC and PRAD over 45 patients showed that the scaling of pipeline processing is linear and perfectly predictable (**Supplementary Table 1** and **Figure 7A**). In addition, **Supplementary Table 1** shows that the entropy pipeline takes a systematically larger time to be completed for high entropy cancer types than for low entropy cancer types. This is also true for the hub diagnosis pipeline (PTTCS). A more careful analysis for 25 patients for LUSC, STAD, LIHC, on one hand, and PRAD, THCA, KIRC, on the other hand, showed that this assumption is statistically significant (**Figure 7B**). Considering the pipeline for hub diagnosis from BLASTx output, we found the time series 50, 94, 137, 187 and 53, 100, 145, 190, for PRAD and NSCLC, respectively. These differences were not significant, but suggest that this pipeline scales similarly to the PTTCS one.

Web Application

The web application implements the graphical interface that allows the user to interact with the forms and their respective accounts (i.e., private areas). As outline above, it is the server that runs Galaxy and hosts MongoDB that stores the up-regulated hubs and patient data introduced by the user, which are necessary to produce the medical record. The frontend includes a succession of forms for data introduction and a private area, which allow access to patient data whenever necessary with user's privileges.

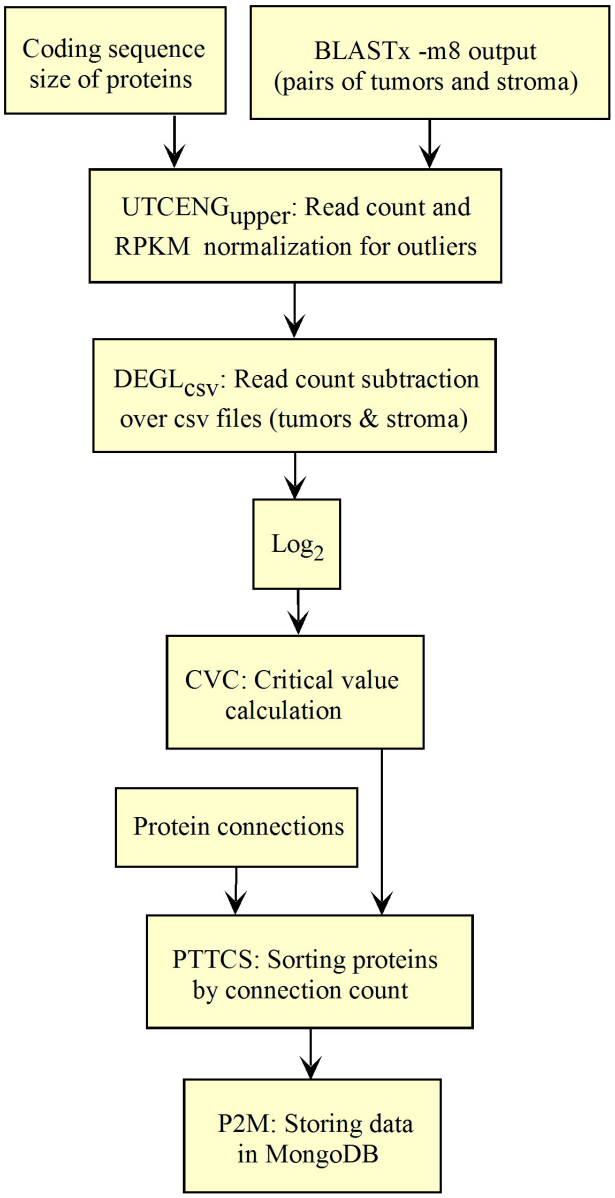

FIGURE 5 | Workflow for collections of BLASTx output processing.

User Private Area

The private area is the section accessed by the user after logging in (see the dashboard in **Figure 8**). The key advantages of a private area are that (i) the user may access their information any time, (ii) sections can be customized, with different levels of privileges, (iii) they can be customized according to business models (Blank and Dorf, 2012).

Dashboard

The dashboard (**Figure 8**) is the first page one sees when accessing the platform after login in from the *welcome* page. On the welcome page, users can register an account. The main goal of the dashboard is gathering all the essential information contained in the portal for the logged in user (e.g., forms to be submitted by users). Thus, users can either introduce the data of their patients or retrieve analysis reports, if they are physicians or administrate the platform, if they are system administrators.

We implemented a simplified version for small devices to fit their screen size and limit the system to the essential (**Figure 8A**). The user is informed when using the system on small devices, which is a benefit compared to Bootstrap. As a result of screen simplification, most of the information from the desktop version (**Figure 8B**) is omitted on small devices, which means that users must access the platform either from desktops or middle size devices (e.g., iPads) for a full-version.

Components

Components in Angular are a set of three types of files: CSS (appearance-related), TS (typescript, coding), and HTML (classic static page). A page is built from at least one component, which can independently interact with each other (Fain and Moiseev, 2018). From a software engineering viewpoint, this technology makes the pages more dynamic and faster, and its parts can be easily reused on other pages. The main components of the dashboard are the menu and central cards. The menu, located upward, displays basic and customized information eventually organized in options. The central cards, movable downward, display information and make them available as active links (e.g., a list of forms submitted by the user).

Protecting Confidential and Sensitive Information

Patients' data are confidential and require protection as stated by policies all over the world (e.g., *Health Insurance Portability and Accountability Act*, HIPAA for the United States). Thus, new users must first register and enter some basic information to gain access to the server.

Login

The *login card* (**Supplementary Figure 5**) is a standard login page. In the current frontend version, we are following a simple login system strategy. Essentially, the user must enter its e-mail and password as previously registered to log in and access the dashboard.

Since we are storing JWT locally, it is up to the user to decide when to log out. Normally, JWT expires after 15 min on a standard basis; we set the expiration time to 1 day. This approach avoids repeated login whenever the JWT expires.

Forms

In the current frontend version, we have two sets of forms: the patient main form (**Supplementary Figure 7**) and the outcome form (**Supplementary Figure 8**). The patient main form is expected to be sent alongside the patient samples, which is independent, while the outcome form is expected to be sent in case of death (for documentation).

All the information related to a patient is stored in different documents and is merged for display using a method called

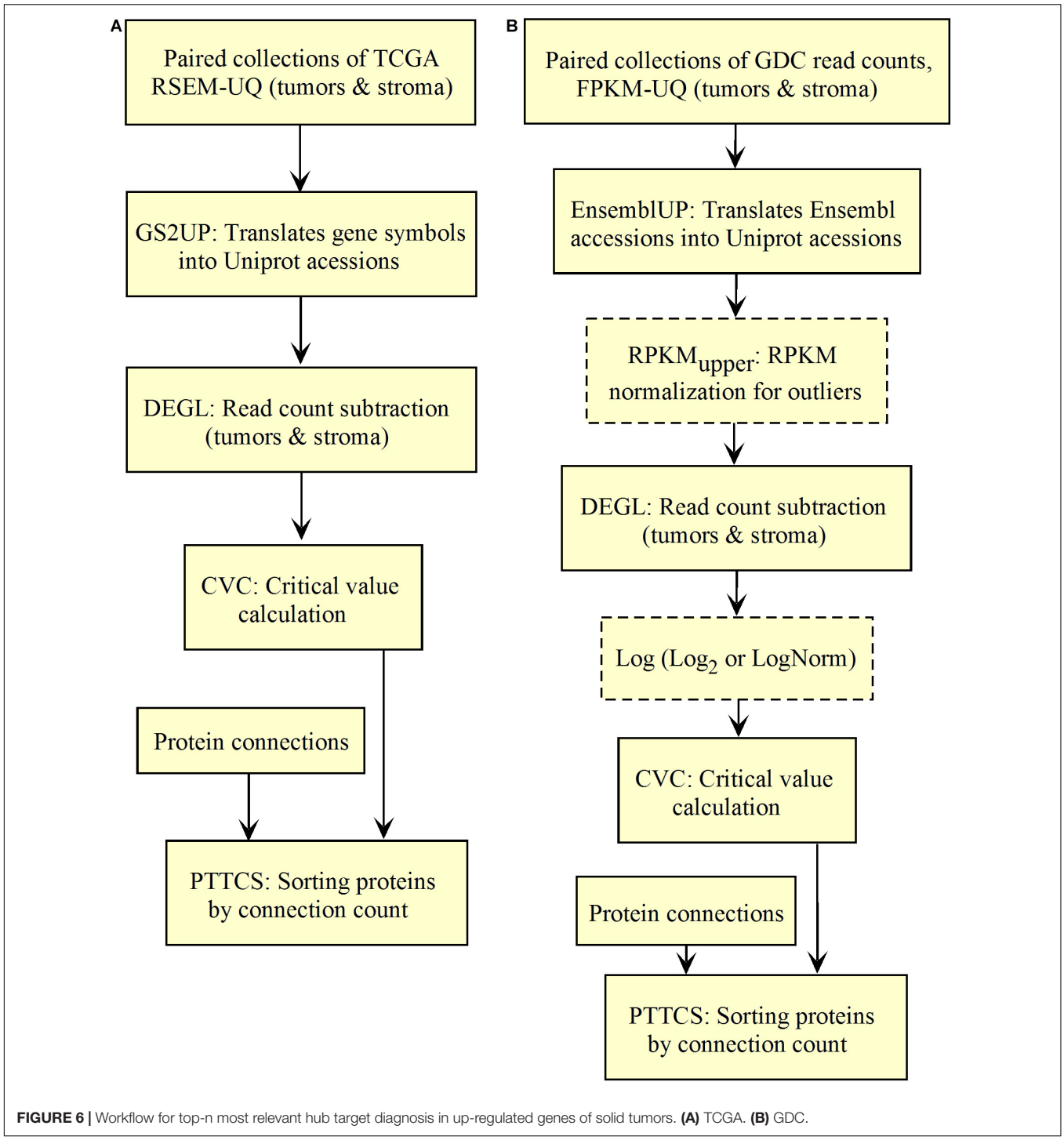

FIGURE 6 | Workflow for top-n most relevant hub target diagnosis in up-regulated genes of solid tumors. **(A)** TCGA. **(B)** GDC.

populate from Mongoose, which enable the information retrieval from other related documents.

Because of this design, we created a *header* form (**Figure 9**), whose function is to (i) collect encrypted patient id, (ii) provide a password for encryption (optional), and (iii) provide privacy-related options.

The form remains in contact with the server for validating information on the background, while the user is filling out the fields; most of the validations are done without communication with the server.

Sensitive information are entered on the first page and encrypted in a similar way to the data introduced through the header. Any form can be recovered from a list of links that are made available on the *movable card* on the dashboard.

Finally, a submission receipt is automatically generated upon form submission (see **Supplementary Figure 9**), which provides

TABLE 6A | Comparative pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to TCGA RSEM-UQ normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av	StDv
HSP90AB1	30.00	80.49	77.78	66.67	56.06	58.06	40.35	83.33	85.71	9	64.27	19.77
YWHAZ	72.00	85.37	93.33	88.89	33.33	29.03	28.07	56.67	36.73	9	58.16	27.26
FN1	26.00	56.10	44.44	77.78	87.88	51.61	85.96	53.33	32.65	9	57.31	22.30
ACTB	26.00	53.66	33.33	44.44	63.64	77.42	38.60	33.33	42.86	9	45.92	16.37
MYH9	14.00	43.90	37.78	33.33	48.48	25.81	43.86	80.00	53.06	9	42.25	18.55
VIM	32.00	12.20	11.11	11.11	93.94	96.77	47.37	23.33	28.57	9	39.60	33.73
RPL10	10.00	46.34	28.89	22.22	80.30	45.16	47.37	10.00	30.61	9	35.66	22.09
EEF1A1	12.00	21.95	24.44	22.22	68.18	22.58	78.95	16.67	22.45	9	32.16	23.93
PKM	NA	92.68	100.00	77.78	77.27	74.19	68.42	70.00	14.29	8	71.83	25.71
HSPA5	60.00	87.80	71.11	77.78	42.42	25.81	NA	43.33	48.98	8	57.15	20.79
HSPB1	NA	41.46	80.00	22.22	42.42	93.55	38.60	40.00	73.47	8	53.97	24.94
HSP90AA1	26.00	65.85	73.33	66.67	NA	48.39	17.54	70.00	57.14	8	53.12	20.97
CLTC	14.00	73.17	24.44	33.33	NA	45.16	12.28	26.67	28.57	8	32.20	19.57
SFN	NA	26.83	71.11	11.11	NA	16.13	NA	10.00	NA	5	27.04	25.52
LRRK2	NA	NA	NA	NA	36.36	54.84	71.93	NA	NA	3	54.38	17.79
VCAM1	NA	NA	NA	11.11	80.30	54.84	NA	NA	NA	3	48.75	35.00
EGLN3	NA	NA	26.67	NA	95.45	NA	NA	NA	NA	2	61.06	48.64
SYNPO	NA	NA	NA	NA	75.76	NA	10.53	NA	NA	2	43.14	46.13

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

TABLE 6B | Comparative pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to GDC $RPKM_{upper}$ normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av	StDv
PKM	16.67	91.23	100.00	87.5	78.873	77.42	80.36	92.59	60.00	9	76.07	25.05
FN1	33.33	70.18	50.00	100	90.141	51.61	87.50	59.26	62.00	9	67.11	21.79
YWHAZ	29.17	63.16	100.00	77.5	50.704	48.39	32.14	70.37	74.00	9	60.60	22.82
UBE2I	10.42	71.93	54.17	47.5	52.113	77.42	83.93	66.67	42.00	9	56.24	22.28
HSP90AB1	54.17	47.37	75.00	57.5	40.845	38.71	19.64	81.48	80.00	9	54.97	20.94
NPM1	60.42	47.37	43.75	50	73.239	35.48	26.79	25.93	58.00	9	46.77	15.79
CTNNB1	27.08	71.93	20.83	27.5	16.901	9.68	69.64	66.67	56.00	9	40.69	25.01
CDKN1A	12.50	52.63	16.67	10	50.704	51.61	71.43	11.11	26.00	9	33.63	23.08
ACTB	NA	36.84	37.50	75	61.972	77.42	33.93	44.44	80.00	8	55.89	19.86
HSP90AA1	27.08	29.82	72.92	70	NA	45.16	10.71	85.19	68.00	8	51.11	26.66
HSPB1	NA	21.05	77.08	40	32.394	77.42	16.07	29.63	64.00	8	44.71	24.70
RPL10	52.08	33.33	18.75	NA	78.873	32.26	28.57	14.81	44.00	8	37.84	20.55
VIM	NA	24.56	NA	12.5	92.958	96.77	35.71	11.11	16.00	7	41.37	37.50
SKP1	12.50	22.81	NA	45	16.901	NA	12.50	11.11	80.00	7	28.69	25.51
TSC22D1	16.67	36.84	NA	NA	12.676	9.68	69.64	NA	12.00	6	26.25	23.45
EGFR	NA	10.53	39.58	NA	74.648	19.35	NA	NA	NA	4	36.03	28.47

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

the user with the information necessary for future access to the forms submitted (see **Supplementary Figure 10**).

DISCUSSION

Galaxy Pipeline

In this report, we presented a workflow for processing RNA-seq data that allows the rational diagnosis of top connected hubs among genes that are up-regulated in tumors according to the non-tumoral peripheral area (stroma). The use of the stroma as a control to measure the malignant differential expression via RNA-seq has been recognized to be equivalent to the use of healthy tissues for this purpose (Finak et al., 2006). Of course, many factors may promote cancer such as chemicals, radiation as well as genetic defects in reparation and replication molecular machinery. To gain inside into such a complex problem as a molecular approach of cancer together with a still-evolving protocol of RNA-seq treatment regarding normalization procedure or error rate (Li et al., 2020), a robust measure was

TABLE 6C | Comparative pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to GDF FPKM-UQ normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av.	StDv
HSP90AB1	81.25	90.91	87.23	86.11	77.46	74.19	64.29	92.59	94.00	9	83.12	9.78
TP53	81.25	75.76	68.09	66.67	78.87	87.10	85.71	66.67	54.00	9	73.79	10.78
TRAF2	45.83	66.67	70.21	83.33	84.51	87.10	35.71	100.00	86.00	9	73.26	20.95
YWHAZ	39.58	81.82	100.00	83.33	66.20	80.65	60.71	66.67	74.00	9	72.55	17.05
PPP1CA	58.33	72.73	78.72	97.22	46.48	61.29	83.93	55.56	74.00	9	69.81	15.84
TRIM27	79.17	90.91	42.55	75.00	52.11	64.52	60.71	66.67	70.00	9	66.85	14.39
FN1	39.58	60.61	42.55	97.22	91.55	54.84	87.50	55.56	32.00	9	62.38	24.08
GRB2	35.42	39.39	36.17	72.22	66.20	93.55	64.29	66.67	66.00	9	59.99	19.39
MAPK6	85.42	69.70	93.62	66.67	45.07	45.16	42.86	37.04	48.00	9	59.28	20.37
GOLGA2	85.42	57.58	59.57	63.89	56.34	67.74	44.64	48.15	46.00	9	58.81	12.75
SNW1	45.83	72.73	68.09	52.78	50.70	58.06	75.00	44.44	58.00	9	58.40	11.28
VCAM1	25.00	72.73	44.68	47.22	88.73	67.74	26.79	66.67	32.00	9	52.40	22.59
CDC37	39.58	27.27	29.79	41.67	70.42	54.84	51.79	74.07	74.00	9	51.49	18.31
MYC	70.83	33.33	63.83	19.44	85.92	77.42	32.14	48.15	30.00	9	51.23	23.95
IKBKE	NA	81.82	76.60	66.67	45.07	80.65	64.29	62.96	58.00	8	67.01	12.44
OTUB1	35.42	69.70	91.49	80.56	21.13	NA	71.43	40.74	60.00	8	58.81	24.24
MDFI	45.83	84.85	87.23	25.00	16.90	NA	80.36	77.78	34.00	8	56.49	29.16
EGFR	27.08	39.39	78.72	NA	95.77	77.42	62.50	44.44	24.00	8	56.17	26.37
HSPB1	NA	45.45	42.55	50.00	40.85	64.52	23.21	37.04	72.00	8	46.95	15.43
YWHAB	22.92	39.39	40.43	77.78	NA	NA	39.29	62.96	72.00	7	50.68	20.30
MAP1LC3B	NA	27.27	NA	27.78	50.704	80.65	60.71	29.63	30.00	7	43.82	20.87
KDM1A	72.92	42.42	36.17	41.67	NA	NA	28.57	40.74	40.00	7	43.21	13.94
WDYHV1	41.67	60.61	65.96	77.78	NA	NA	NA	33.33	66.00	6	57.56	16.73
KSR1	NA	30.30	19.15		84.507	NA	NA	18.52	20.00	5	34.50	28.37

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

needed. We found this measure in the degree entropy. Entropy offers the benefit to be independent of sample size. In this report, we calibrated our approach by reference to OS, but after an optimization round for the treatment of RNA-seq data, other factors could be taken into account to understand how they interact with the signaling network complexity.

Normalization of raw read counts account for (i) within-sample effects induced by factors such as coding sequence size (Oshlack and Wakefield, 2009), GC-content (Risso et al., 2011), (ii) between-sample effects such as sequencing depth (total number of molecules sequenced) (Robinson and Oshlack, 2010), and (iii) batch effect (Tom et al., 2017). As underlined by Evans et al. (2018), "normalization methods perform poorly when their assumptions are violated." Thus, the exercise is to "select a normalization method with assumptions that are met and that produces a meaningful measure of expression for the given experiment."

Following these recommendations, we must first consider that the purpose of our approach is to list the top-n most relevant target among subnetworks of genes that are up-regulated in tumor samples compared to their controls. Consequently, the complexity of the up-regulated gene subnetwork must be coherent with the 5-years OS. Indeed, our supporting hypothesis is that the complexity of the malignant subnetwork or the number of times that the malignant subnetwork can reorganize itself after perturbation is in line with its information content, i.e., its Shannon entropy. This is the reason why it makes sense to optimize the normalization process for maximizing the coefficient of correlation between entropy and 5-years OS. We aimed to diagnose the subnetwork complexity because it is correlated to the 5-years OS and this is important for therapy's success (whatever being performed with drugs or biopharmaceuticals) in the context of a personalized approach of oncology.

The PDF and CDF functions of the Python's scipy package allowed the calculation of the critical values given the density of probability of non-differentially expressed genes. These distribution are rather similar regardless of the RNA-seq considered for a given normalization process. These genes are thousands while the up-regulated ones are hundreds, which makes critical value determined in this way rather precise and reproducible. Concerning the statistical significance of the method we applied, one has to say that we face a classification problem. In such circumstances, one usually looks for the optimization between false positive and false negative rates. However, when dealing with medical purpose, one has to look to bias the classification process toward the minimization of false-positive rate to reduce toxic drug collateral effects to patients that would derive from hubs still expressed at a significant level in the stroma (this consideration does not concern drug toxicity due to off-target effects). There is a compromise between minimizing the false positive rate and the availability of hub targets for

TABLE 6D | Pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to successive processing through $RPKM_{upper}$ and LogNorm normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av	StDv
HSP90AB1	83.33	73.68	87.50	65.00	76.06	74.19	42.86	96.30	98.00	9	77.44	16.92
TP53	79.17	84.21	66.67	65.00	74.65	83.87	73.21	77.78	66.00	9	74.51	7.42
YWHAZ	47.92	70.18	100.00	87.50	63.38	67.74	46.43	77.78	86.00	9	71.88	17.97
TRAF2	47.92	47.37	72.92	85.00	84.51	87.10	26.79	96.30	86.00	9	70.43	23.85
FN1	43.75	70.18	52.08	100.00	90.14	54.84	87.50	59.26	62.00	9	68.86	19.43
PPP1CA	56.25	68.42	79.17	95.00	46.48	51.61	67.86	70.37	76.00	9	67.91	14.97
GRB2	43.75	43.86	39.58	87.50	71.83	87.10	46.43	81.48	94.00	9	66.17	22.45
MAPK6	85.42	66.67	95.83	65.00	49.30	58.06	42.86	55.56	72.00	9	65.63	16.92
GOLGA2	87.50	61.40	70.83	65.00	54.93	51.61	37.50	62.96	80.00	9	63.53	15.00
TRIM27	79.17	68.42	56.25	62.50	54.93	45.16	53.57	77.78	68.00	9	62.86	11.47
SNW1	39.58	54.39	72.92	47.50	43.66	45.16	44.64	74.07	82.00	9	55.99	15.94
HSCB	56.25	56.14	68.75	47.50	74.65	29.03	51.79	59.26	58.00	9	55.71	12.95
VCAM1	31.25	47.37	47.92	45.00	87.32	67.74	30.36	81.48	46.00	9	53.83	20.48
CDC5L	43.75	43.86	68.75	52.50	53.52	48.39	16.07	77.78	78.00	9	53.62	19.49
MYC	72.92	40.35	66.67	10.00	85.92	77.42	30.36	55.56	36.00	9	52.80	25.20
OTUB1	37.50	54.39	83.33	77.50	15.49	22.58	62.50	40.74	54.00	9	49.78	23.03
IKBKE	10.42	75.44	60.42	72.50	32.39	64.52	76.79	22.22	22.00	9	48.52	26.46
REL	35.42	29.82	45.83	52.50	26.76	32.26	25.00	70.37	16.00	9	37.11	16.58
EGFR	35.42	57.89	79.17	NA	95.77	67.74	53.57	51.85	44.00	8	60.68	19.55
MDFI	47.92	80.70	87.50	30.00	15.49	NA	82.14	81.48	26.00	8	56.40	29.80
GABARAPL2	37.50	22.81	NA	20.00	11.27	25.81	30.36	59.26	80.00	8	35.87	22.86
YWHAB	20.83	31.58	22.92	75.00	NA	16.13	32.14	22.22	28.00	8	31.10	18.57
LRRK2	NA	49.12	NA	15.00	83.10	87.10	89.29	22.22	14.00	7	51.40	34.87
LNX1	50.00	61.40	66.67	42.50	NA	NA	48.21	29.63	44.00	7	48.92	12.32
MAP1LC3B	NA	35.09	NA	17.50	42.25	80.65	46.43	62.96	52.00	7	48.13	20.16

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

therapy. A larger p-value ($p > 0.025$) would release a larger list of up-regulated genes with more hub targets; a larger list of potential drugs for the case under consideration, but also a larger probability of toxic effects on the stroma. In contrast, lower p-value ($p < 0.025$) will minimize toxic effect of therapy to patient, but would also decrease the number of potential hubs for therapy. Of course, this consideration neglects the tissue specific expression of genes and a gene that is up-regulated in a tumor compared to its stroma could also be up-regulated in another tissue, on a normal basis. Here, we neglected this issue, but it is possible to preferentially target tumors through nanoparticle therapy or by local application.

As pointed out by Abbas-Aghababazadeh et al. (2018), it is possible that some of the estimated *latent factors* are not technical artifacts but rather represent true biological features reflected in the data. The correction of these latent factors may introduce unwanted biases. Here, we did not want to stabilize the subnetwork size variance (Smyth, 2004; Cloonan et al., 2008; Love et al., 2014; Holmes and Huber, 2019) because we believe that it is part of the challenge. One cannot exclude the possibility of network size varying among samples according to the specificities of genome deregulation proper to a given tumor. Despite commonalities that were recognized between tumors of the same cancer type, many features such as gene demethylation, copy numbers, somatic crossing over, and chromosome karyotype contribute to the specificity of the molecular phenotype of a tumor and it is the correct diagnosis of these specificities that can make the difference in terms of patient benefits (Duesberg et al., 2005; Ozery-Flato et al., 2011; Ogino et al., 2012; Grade et al., 2015; Bloomfield and Duesberg, 2016; Ye et al., 2018; Xia et al., 2019).

According to the considerations just outlined, the size of the malignant subnetwork is also important because it directly affects the number of targets available for therapy. The size of the malignant subnetworks also depends on the normalization process. There is a tradeoff between the size of the malignant subnetwork and the level of tumor personalization that is effectively reported by the top-n targets as a result of the normalization process. From our perspective, the normalization corresponding to GDC FPKM-UQ and $RPKM_{upper}$ + LogNorm generate subnetworks that are too large since they represent as much as 20% of the human proteome (>4,000 genes). By contrast, subnetworks produced by GDC $RPKM_{upper}$ + Log2 normalization account for between 2 and 5% of the human proteome, which seems to be more realistic (Danielsson et al., 2013; Malvia et al., 2019).

The target lists that we found with the various normalization methods presented here were consistent among one another and

TABLE 6E | Pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to successive processing through $RPKM_{upper}$ and Log2 normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av	StDv
HSP90AB1	77.08	66.67	85.42	94.00	64.79	64.52	35.71	96.30	94.00	9	75.39	19.73
YWHAZ	41.67	70.18	100.00	82.00	60.56	67.74	46.43	74.07	82.00	9	69.41	18.21
TP53	72.92	70.18	64.58	60.00	66.20	74.19	58.93	70.37	60.00	9	66.37	5.86
FN1	41.67	70.18	52.08	62.00	90.14	54.84	87.50	59.26	62.00	9	64.41	15.92
NPM1	72.92	68.42	72.92	60.00	85.92	41.94	51.79	59.26	60.00	9	63.68	12.98
YWHAG	52.08	40.35	87.50	68.00	52.11	80.65	10.71	59.26	68.00	9	57.63	22.91
CDC37	12.50	28.07	25.00	86.00	67.61	38.71	44.64	81.48	86.00	9	52.22	28.54
MAPK6	79.17	43.86	87.50	54.00	22.54	29.03	23.21	44.44	54.00	9	48.64	23.06
MYH9	39.58	40.35	37.50	62.00	43.66	12.90	23.21	81.48	62.00	9	44.74	20.98
PKM	14.58	73.68	41.67	22.00	47.89	32.26	78.57	51.85	22.00	9	42.72	22.68
HSPB1	NA	49.12	89.58	92.00	59.15	93.55	50.00	37.04	92.00	8	70.31	23.74
RPL10	66.67	56.14	45.83	74.00	88.73	54.84	44.64	NA	74.00	8	63.11	15.41
OTUB1	27.08	42.11	83.33	78.00	NA	16.13	48.21	66.67	78.00	8	54.94	25.35
YWHAB	14.58	35.09	39.58	84.00	NA	35.48	26.79	77.78	84.00	8	49.66	27.82
YBX1	31.25	21.05	66.67	56.00	49.30	41.94	NA	70.37	56.00	8	49.07	16.96
MYC	66.67	15.79	56.25	28.00	73.24	67.74	NA	40.74	28.00	8	47.05	21.80
EGFR	16.67	40.35	68.75	28.00	95.77	58.06	30.36	NA	28.00	8	45.75	26.53
CSNK2A1	20.83	21.05	85.42	30.00	11.27	41.94	NA	51.85	30.00	8	36.54	23.49
GRB2	NA	15.79	16.67	64.00	32.39	87.10	NA	48.15	64.00	7	46.87	26.75
TUBA1A	NA	47.37	10.42	18.00	59.15	54.84	71.43	NA	18.00	7	39.89	24.06
LRRK2	NA	38.60	NA	NA	57.75	67.74	76.79	NA	NA	4	60.22	16.38
VCAM1	NA	10.53	10.42	NA	84.51	61.29	NA	NA	NA	4	41.69	37.27
LZTS2	NA	36.84	NA	NA	NA	29.03	71.43	NA	NA	3	45.77	22.56
EGLN3	NA	NA	22.92	NA	83.10	NA	NA	NA	NA	2	53.01	42.56

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

with that of Conforte et al. (2019). The normalization method corresponding to the best compromise according to subnetwork size, correlation, and the target list was $RPKM_{upper}$ + Log2, and it is that method that was, therefore, kept for new sample analyses.

To be coherent with former studies, we included LUAD and BRCA, however, these two cancer types discredited the analyses for two obvious reasons: (i) In the case of LUAD, the samples of *raw counts* did not match those of FPKM-UQ, which prohibit direct comparison between both datasets and raised the question why FPKM-UQ normalization was not performed on a large proportion of *raw counts* files and why, on the other hand, other samples were taken into account in the FPKM-UQ processing. This discrepancy may explain why LUAD does not match the regression line in GDC $RPKM_{upper}$ + Log2, while it does in GDC FPKM-UQ; (ii) In the case of BRCA, after filtrating samples for matching between *raw counts* and FPKM-UQ, the total sample size was less than 20, which is not sufficient for statistical significance given subtype heterogeneity. BRCA is composed of four subtypes whose 5-years OS varies between 70 and 82%. **Figure 3** shows that depending on sampling, BRCA could very well match the linear regression.

The relevance of inhibiting hub of connections has been proven mathematically by Albert et al. (2000) and its benefit for patients has been confirmed by Conforte et al. (2019) through Shannon entropy analysis. The negative correlation found between the subnetwork entropy and the 5-years OS is in agreement with the results obtained later on from the modeling of basins of attraction in BC with Hopfield network (Conforte et al., 2020). This study revealed that five tumor samples converged toward the basin of attraction associated with control samples instead of the tumor ones. Those samples were associated with a good prognosis, initial stages of tumor development, and four of them presented the smallest subnetwork entropy among the dataset of 70 tumor samples under study.

As the research concept has been validated through different approaches, the workflow presented here was built with the aim of automating the analysis, which will allow its translation to the medical context. With that concern, the larger time needed for entropy and PTTCS pipelines to be completed when analyzing high entropy cancer types compared to the processing time spent with low entropy cancer types suggests a positive relationship between subnetwork complexity and their processing time. If confirmed, this observation means that the computation model, presented here, reproduces a main biological feature of cancer that is the larger complexity associated with subnetwork of up-regulated genes in aggressive tumors. In any case, the difference in the processing time of the PTTCS pipeline for high and low entropy cancers was not large (∼50 s for 25 patients).

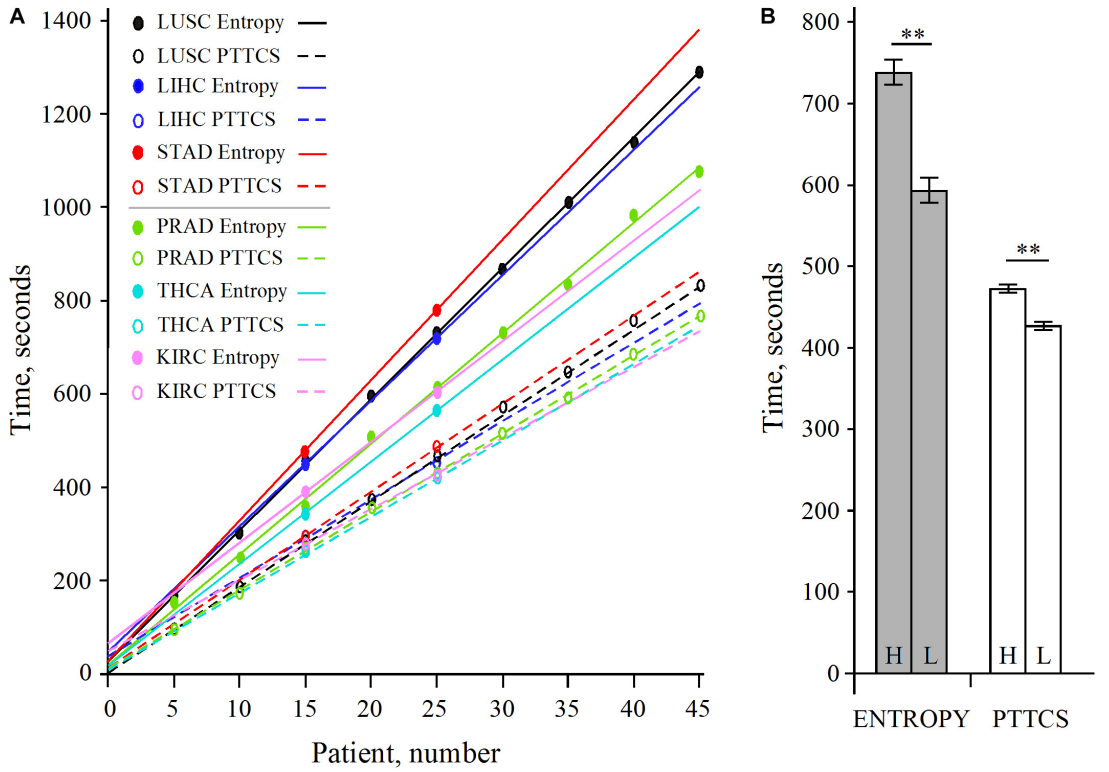

FIGURE 7 | Scaling of pipeline from **Figure 2B** (entropy) and **Figure 6B** (PTTCS) using GDC read counts (see data in **Supplementary Table 1**). **(A)** Linear scaling for 5–45 patients in LUSC and PRAD as well as 15 and 25 patients for STAD, LIHC, THCA, and KIRC. **(B)** Statistical analysis of scaling for high entropy (H) cancer (LUSC, LIHC, STAD) and low entropy (L) cancer (KIRC, PRAD, THCA) for entropy (gray) and PTTCS (white) pipelines. **Significant at $\alpha \leq 0.01$ for $k = 4$ degrees of freedom. The horizontal bars are for the standard error of the mean (SEM).

We believe that our strategy will contribute constructively to cancer treatment because the molecular phenotype of a cell is directly connected to its genetic alterations, which is not necessarily the case for genomic alterations. Genomic alterations allow a diagnosis based on probabilistic data obtained with large patient cohorts. By contrast, the molecular phenotype portraits the cell or the genomic disease and points to proteins that should be targeted in the first instance to disrupt malignant phenotypes while affecting the healthy one the least possible.

The phenotype approach also reflects which genes that malign cells most need to maintain themselves in the tissue given its selective constraints. In any pathogenic relationship, one distinguishes between *primary* and *secondary determinants* of the disease (Yoder, 1980). The primary determinants are those that make the relationship compatible (qualitative) and the secondary determinants are those that deal with its quantitative expression (virulence). Thus, the question to deal with, in the case of cancer, is to target primary determinants. When considering gene expression, one may reason that the heterogeneity is something related to secondary determinants (it is not because a cell is mutating that the new mutations are worse than the previous ones). Actually, it has been well described that a tumor developed by the accumulation of mutations in a small number of key oncogenes or suppressor genes in stem cells and that the probability of this event to occur is very low (Hornsby et al., 2007; Belikov, 2017). Thus, there is a difference between these primary mutations that allow the tumor to establish itself and the secondary ones that may affect its aggressiveness. On the same line, when one sequences the mRNA of a tumor area, one takes the gene expression profile of many cells into account. By consequence, secondary mutations promoting or inhibiting a given gene in different cell lineages inside the same tumor compensate themselves. By contrast, those genes that are key to maintain a malignant cell line will be positively selected to remain up-regulated in most cells and, therefore, if one detects a gene that is up-regulated in a tumor by comparing its expression level with the surrounding stroma, it means that it is essential for malign cell survival.

Considering the number of hubs to target, the results obtained by Conforte et al. (2019) suggest 3–10, on average. Other authors already suggested such complex mixes (Calzolari et al., 2007, 2008; Preissner et al., 2012; Hu et al., 2016; Antolin et al., 2016; Lu et al., 2017). Three to ten specific drugs may appear a small number to control such a complex disease as cancer, but the cell death induction may be explained by a cascading effect, which is larger when targeting hubs as suggested before (Carels et al., 2015a; Barabási, 2016; Tilli et al., 2016; Conforte et al., 2019). According to Conforte et al. (2019), this cascading effect would be inversely proportional to the tumor aggressiveness. The pitfall is that the number of specific drugs for hub targets

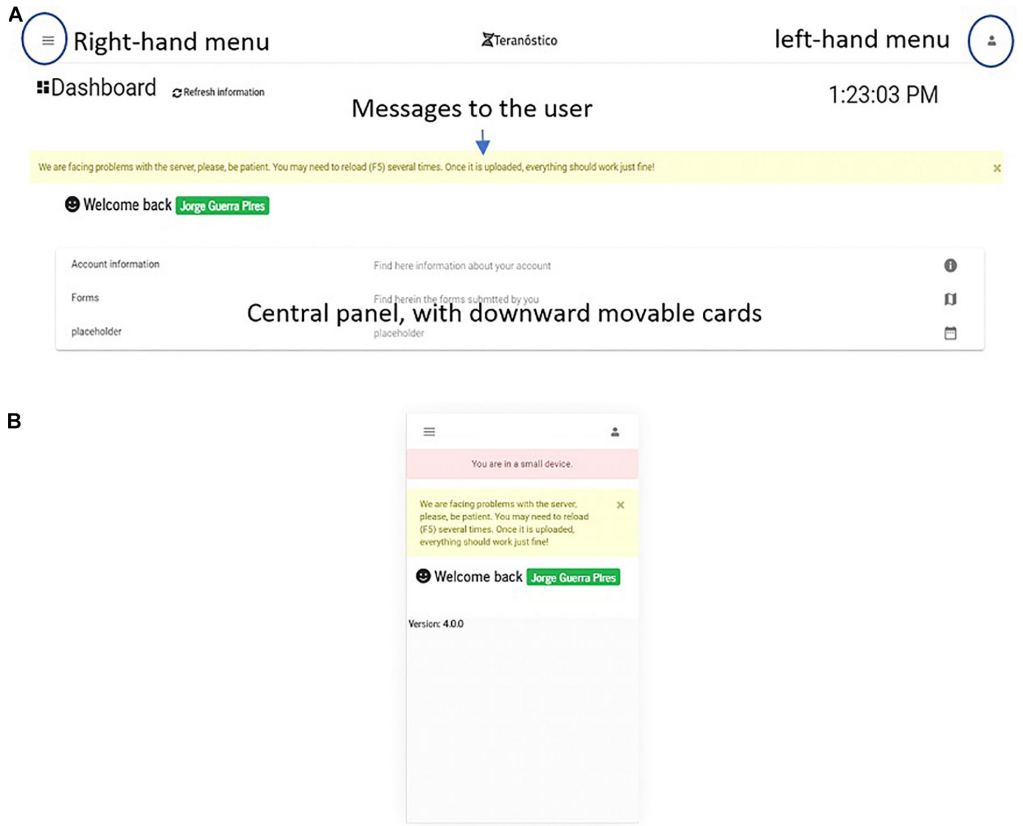

FIGURE 8 | Dashboard on **(A)** a desktop and **(B)** an extra small devices (Galaxy S5).

that are approved by FDA is still very small (Antolin et al., 2016). While new drugs and biopharmaceuticals or products of other strategies continuously appear, key targets remain the same. Some are highly personalized and often secondary while others are constant across tumor types or within a tumor type; these last targets play, in most likelihood, a primary role in the disease and it is essential to diagnose them (even if only for their prognostic value). In addition, nothing prohibits the combination of specific drugs with cytotoxic or hormonal treatments (Nikanjam et al., 2016). The idea is to improve as much as possible the rational drug use to maximize the patient benefit. Many patients are dying from the toxic collateral effect of the chemotherapy; it would be a great success if the use of specific drugs in a standard therapy protocol could enable to decrease the dose of cytotoxic drugs and improve the therapy acceptance by patients in some specific cases in the context of theranostics. For this kind of exercise, an automated pipeline is needed and a clinical trial testing the validity of hubs as potential molecular targets is urgent.

The replication number that can be done for RNA-seq is another limitation given the still high cost of this technology. Thus, analyses as the one described in our manuscript are expected to be done only once per time in a time series for each patient. According to Barabasi's theory (Barabási, 2016), hubs with the same connection rate are expected to have the same disarticulation effect on the signaling network. On a clinical basis, *p*-values (here critical value) may be adapted to the specific case of each patient. On the same line of reasoning, our methodology can be easily adapted taking into account more powerful bioinformatics tools and statistical analysis, but this issue is beyond the scope of this report. For such a methodology improvement, we believe that entropy is a good measure because it is universal, robust, and not dependent on sample size. Different combinations of normalization and statistical analyses as those reported by Li et al. (2020) can be compared in the same framework we presented here and in Conforte et al. (2019), by looking at how they may maximize the correlation coefficient of the negative relationship between entropy and OS. Of course, this depends on accepting the hypothesis that more aggressive tumors have more complex signaling networks, but again, this statement has been repeatedly claimed by several authors worldwide and along several years (Teschendorff and Severini, 2010; van Wieringen and van der Vaart, 2011; Breitkreutz et al., 2012; West et al., 2012; Banerji et al., 2015). If this hypothesis is true, the negative correlation between entropy and OS may serve as a calibration to study the optimization of RNA-seq methodologies and the influence of other factors in cancer development and dynamics.

Cancer is a genomic disease that affects DNA replication checkpoints through mutations of key oncogenes and suppressor genes (Lee and Muller, 2010). There are ten main hallmarks

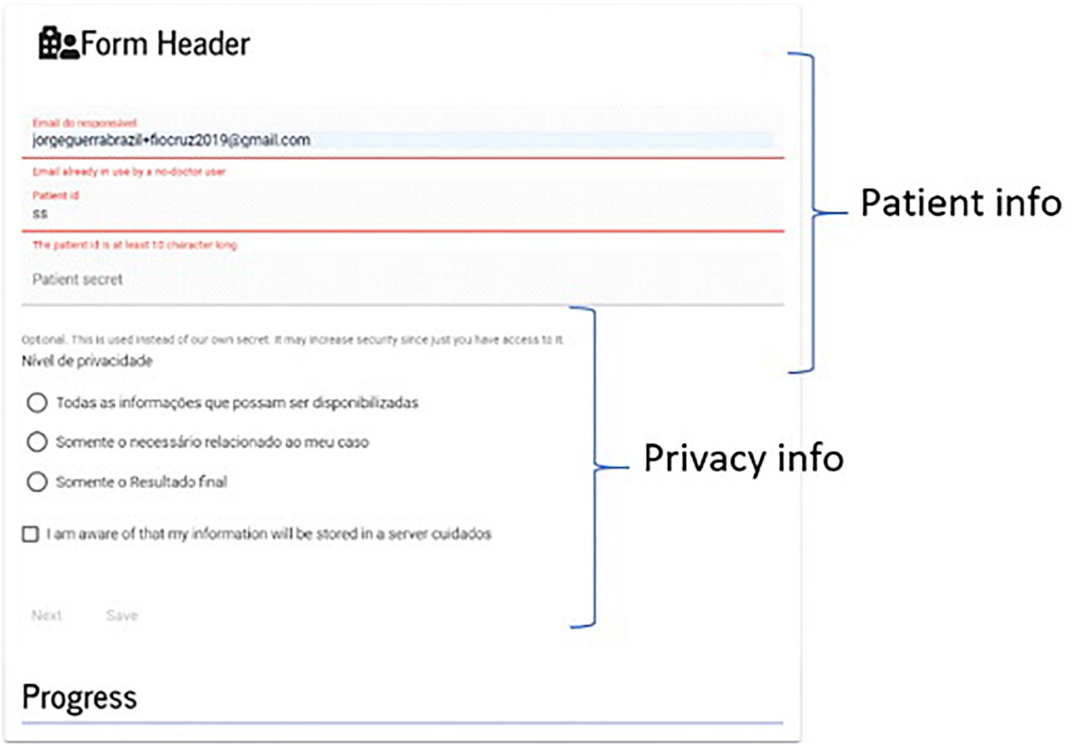

FIGURE 9 | Form header. A user is warned when leaving without saving the form. The patient password is encrypted and kept on the server using a specialized type of file; nonetheless, users can choose their own passwords.

for cancer from which uncontrolled division is the key one (Hanahan and Weinberg, 2011). When the disease is taken at a late stage, it may have spread in the body through metastasis and secondary tumors may have different molecular profiles. In such late tumor stages, an approach of cancer therapy only based on personalized oncology would in most likelihood be unsuccessful (Ashdown et al., 2015). However, specific drugs could increase the patient benefit by supplementing traditional therapies based on cytotoxic drugs. As a consequence, the maximum benefit of a personalized oncology approach of solid tumor therapy based on a molecular phenotype diagnosis is in the early stages of malign cell multiplication. Despite its limitations, the phenotype approach of molecular diagnosis proposed here is needed for rational drug (or biopharmaceutical) therapy to maximize patient benefit.

At the moment, the methodology and the web site that we described here can be assimilated to *laboratory developed tests* (LDT). It is notorious that LDT for being a type of *in vitro* diagnostic test designed, manufactured, and used within a single laboratory is poorly supported by oncologists (8%) and pathologists (12%) because of the legitimate fair of innovation. Biomarkers and CDs strongly depend on the regulation by official organizations for their acceptance by health decision-makers (Novartis, 2020). However, barriers by regulation are no reason to stop the innovation necessary for progress. Otherwise, regulation fails with its purpose of protecting lives (see Carels et al., 2020 for a review).

Web Application

System biology has gained considerable attention in medical sciences in the last decade thanks to the ever-increasing computer power. However, system biology models can be tricky to use or to interpret by non-experts in modeling. A recurrent question is how to integrate models into the physician daily lives such that they could best participate in their decision-making process. One potential solution, which seems to be the predominant one on the current state of the art, is by packing algorithms into software bundles and to make them available by user-friendly interfaces, such that little, or even no, expertise is required to use them. This is the paradigm we followed in this report.

The power and diversity of Angular programmed with TypeScript enable to expand the functionalities of the prototype proposed here in future versions, including the implementation of heavy calculations on the frontend side.

We chose MongoDB for storing genetic and medical records even if Galaxy has its own database system (postgreSQL). Our choice of MongoDB was motivated by the care of keeping coherence with MEAN stack, and also because of the power of MongoDB for Big Data storage. In addition, MongoDB is a non-relational database (NoSQL), which allows the storage of data in different formats within the same database.

Our implementation of online forms offers the possibility of creating new functions such as data validation. Data can be validated by comparing frontend to backend information through the database and making sure, for instance, that an

entered e-mail does not already belong to someone else already registered in the system.

Finally, one common concern on web-programming is to minimize client communication with the server to maximize performance. For such purpose, we implemented a process of form validation on the frontend side. Since we are using FormBuilder (see for more details Fain and Moiseev, 2018), there are a set of built-in validation routines, and the possibility to easily create customized validation, thus any specific demand concerning data validation can be handled on future versions using the current source code.

CONCLUSION

In a successive set of publications, we developed a rational methodology for the diagnosis of connection hubs among up-regulated genes of malignant subnetworks. This strategy is an application of graph theory, whose relevance has been mathematically proven by Albert et al. (2000). The inference of this theory into biological systems performed by Carels et al. (2015a) has been successfully validated on malignant cells by Tilli et al. (2016) and extended to tumor tissues by Conforte et al. (2019).

Here, in a translational oncology effort, we outlined a workflow that automated that research and allows its application to a large set of RNA-seq data to interact with public entities of the oncological sector, such as pharmaceutical companies, hospitals, diagnostic laboratories, public health care systems, and insurance groups around the world.

We belief that innovation in new translational solutions, like the one outlined here, is an imperative attribute of research centers; however, other agents such as (i) pharmaceutical companies may certainly help these initiatives with their experience concerning regulation, market barriers, financial support and (ii) startups whose processing speed and innovation potential were already well-documented (Blank and Dorf, 2012).

Herein, we aimed at transcending basic cancer inferences to bring a solution for clinical applications on a global scale.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://galaxy.cdts.fiocruz.br/.

AUTHOR CONTRIBUTIONS

JP contributed for the development of the web application and wrote the corresponding sections. GS built the galaxy environment. TW wrote the Phython script. AC did the R analysis and contribute to the pipeline logic. DP prepared the medical forms. FS contributed with manuscript writing. NC wrote the Perl scripts, set the pipeline logic up, and wrote and managed the manuscript writing. All the authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by a grant (E-26/290.077/2017 - 227190) to NC and a fellowship (E-26/260.046/2019 - 242550) to JP from Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.624259/full#supplementary-material

Supplementary Figure 1 | Adaptative display according to device screen size (source: Fain and Moiseev, 2018).

Supplementary Figure 2 | Flowchart of main form filling for exam request.

Supplementary Figure 3 | Flowchart of outcome form filling.

Supplementary Figure 4 | Authentication process. A *Guard* function double-check a user's requisition and if access conditions are met the user is allowed to see the content of a requested page (green arrows). By contrast, if something went wrong (e.g., token expired), the access is denied (red arrow).

Supplementary Figure 5 | Login card.

Supplementary Figure 6 | Example of main form options.

Supplementary Figure 7 | Example of outcome form being implemented.

Supplementary Figure 8 | The page #1 of the main form given as an example.

Supplementary Figure 9 | Receipt of main form submission.

Supplementary Figure 10 | Form after retrieval from the Dashboard (the entire form does not fit the page).

Supplementary Table 1 | Scaling of pipeline from **Figure 2B** (entropy) and **Figure 6B** (PTTCS) using GDC read counts (see **Figure 7**).

REFERENCES

Abbas-Aghababazadeh, F., Li, Q., and Fridley, B. L. (2018). Comparison of normalization approaches for gene expression studies completed with highthroughput sequencing. *PLoS One* 13:e0206312. doi: 10.1371/journal.pone.0206312

Afgan, E., Baker, D., Batut, B., Beek, M., Bouvier, D., Čech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* 406, 378–382. doi: 10.1038/35019019

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106.

Antolin, A. A., Workman, P., Mestres, J., and Al-Lazikani, B. (2016). Polypharmacology in precision oncology: current applications and future prospects. *Curr. Pharm. Des.* 22, 6935–6945. doi: 10.2174/1381612822666160923115828

Ashdown, M. L., Robinson, A. P., Yatomi-Clarke, S. L., Ashdown, M. L., Allison, A., Abbott, D., et al. (2015). Chemotherapy for late-stage cancer patients: meta-analysis of complete response rates. *F1000Res* 4:232. doi: 10.12688/f1000research.6760.1

Awazu, A., Tanabe, T., Kamitani, M., Tezuka, A., and Nagano, A. J. (2018). Broad distribution spectrum from gaussian to power law appears in stochastic

variations in RNA-seq data. *Sci. Rep.* 8:8339. doi: 10.1038/s41598-018-26735-4

Balwierz, P. J., Carninci, P., Daub, C. O., Kawai, J., Hayashizaki, Y., Van Belle, W., et al. (2009). Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol.* 10:R79.

Banerji, C. R. S., Severini, S., Caldas, C., and Teschendorff, A. E. (2015). Intratumour signalling entropy determines clinical outcome in breast and lung cancer. *PLoS Comput. Biol.* 11:e1004115. doi: 10.1371/journal.pcbi.1004115

Barabási, A.-L. (2016). *Network Science*. Cambridge: Cambridge University Press, 475.

Belikov, A. V. (2017). The number of key carcinogenic events can be predicted from cancer incidence. *Sci. Rep.* 7:12170. doi: 10.1038/s41598-017-12448-7

Blank, S., and Dorf, B. (2012). *The Startup Owner's Manual: The Step-By-Step Guide for Building a Great Company*. Bartlett: K & S Ranch.

Bloomfield, M., and Duesberg, P. (2016). Inherent variability of cancer-specific aneuploidy generates metastases. *Mol. Cytogenet.* 9:90. doi: 10.1186/s13039-016-0297-x

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for highdensity oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi: 10.1093/bioinformatics/19.2.185

Bradshaw, S., Brazil, E., and Chodorow, K. (2019). *MongoDB: The Definitive Guide: Powerful and Scalable Data Storage*, 3rd Edn. Newton, MA: O'Reilly Media, Inc, 514.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Breitkreutz, D., Hlatky, L., Rietman, E., and Tuszynski, J. A. (2012). Molecular signaling network complexity is correlated with cancer patient survivability. *Proc. Natl. Acad. Sci. U S A.* 109, 9209–9212. doi: 10.1073/pnas.1201416109

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

Calzolari, D., Bruschi, S., Coquin, L., Schofield, J., Feala, J. D., Reed, J. C., et al. (2008). Search algorithms as a framework for the optimization of drug combinations. *PLoS Comput. Biol.* 4:e1000249. doi: 10.1371/journal.pcbi.1000249

Calzolari, D., Paternostro, G., Harrington, P. L., Piermarocchi, C., and Duxbury, P. M. (2007). Selective control of the apoptosis signaling network in heterogeneous cell populations. *PLoS One* 2:e547. doi: 10.1371/journal.pone.0000547

Campbell, P. J., Getz, G., Korbel, J. O., and The Icgc/Tcga Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature* 578, 82–93. doi: 10.1038/s41586-020-1969-6

Carels, N., Conforte, A. J., Lma, C. R., and da Silva, F. A. B. (2020). "Challenges for the optimization of drug therapy in the treatment of cancer," in *Computational Biology*, 1ed Edn, Vol. 32, eds F. A. B. da Silva, N. Carels, T. M. dos Santos, and F. J. P. Lopes (Cham: Springer International Publishing), 163–198. doi: 10.1007/978-3-030-51862-2_8

Carels, N., Tilli, T., and Tuszynski, J. A. (2015a). A computational strategy to select optimized protein targets for drug development toward the control of cancer diseases. *PLoS One* 10:e0115054. doi: 10.1371/journal.pone.0115054

Carels, N., Tilli, T. M., and Tuszynki, J. A. (2015b). Optimization of combination chemotherapy based on the calculation of network entropy for protein-protein interactions in breast cancer cell lines. *EPJ Nonlinear Biomed. Phys.* 3:6.

Catharina, L., de Menezes, M. A., and Carels, N. (2018). "System biology to access target relevance in the research and development of molecular inhibitors," in *Theoretical and Applied Aspects of System Biology. Computational Biology*, 1ed Edn, Vol. 27, eds F. A. B. da Silva, N. Carels, and F. Paes Silva Jr. (Cham: Springer International Publishing), 221–242. doi: 10.1007/978-3-319-74974-7_12

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619. doi: 10.1038/nmeth.1223

Collins, F. S., and Varmus, H. (2015). A new initiative on precision medicine. *N. Engl. J. Med.* 372, 793–795. doi: 10.1056/nejmp1500523

Conforte, A. J., Alves, L. D., Coelho, F. C., Carels, N., and da Silva, F. A. B. (2020). Modeling basins of attraction for breast cancer using Hopfield networks. *Front. Genet.* 11:314. doi: 10.3389/fgene.2020.00314

Conforte, A. J., Tuszynski, J. A., da Silva, F. A. B., and Carels, N. (2019). Signaling complexity measured by shannon entropy and its application in personalized medicine. *Front. Genet.* 10:930. doi: 10.3389/fgene.2019.00930

Dagnelie, P. (1970). *Théorie et méthodes Statistiques: Applications Agronomiques Vol. 2. Les méthodes de l'inférence Statistique*. Gembloux: J. Duculot, 451.

Danielsson, F., Skogs, M., Huss, M., Rexhepaj, E., O'Hurley, G., Klevebring, D., et al. (2013). Majority of differentially expressed genes aredown-regulated during malignant transformationin a four-stage model. *Proc. Natl. Acad. Sci. U S A.* 110, 6853–6858. doi: 10.1073/pnas.1216436110

Deelman, E., Gannon, D., Shields, M., and Taylor, I. (2009). Workflows and e-Science: an overview of workflow system features and capabilities. *Future Generat. Comp. Systems* 25, 528–540. doi: 10.1016/j.future.2008.06.012

Duesberg, P., Li, R., Fabarius, A., and Hehlmann, R. (2005). Aneuploidy and cancer: from correlation to causation. *Cell. Oncol.* 27, 293–318.

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* 19, 776–792. doi: 10.1093/bib/bbx008

Fain, Y., and Moiseev, A. (2018). *Angular Development with TypeScript, Second Edition*. Shelter Island, NY: Manning Publications, 560.

Finak, G., Sadekova, S., Pepin, F., Hallett, M., Meterissian, S., Halwani, F., et al. (2006). Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. *Breast Cancer Res.* 8:R58.

Grade, M., Difilippantonio, M. J., and Camps, J. (2015). Patterns of chromosomal aberrations in solid tumors. *Recent Results Cancer Res.* 200, 115–142. doi: 10.1007/978-3-319-20291-4_6

Guo, X. E., Ngo, B., Modrek, A. S., and Lee, W.-H. (2014). Targeting tumor suppressor networks for cancer therapeutics. *Curr. Drug Targets* 15, 2–16. doi: 10.2174/1389450114666140106095151

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Holmes, S., and Herber, C. (2019). *Getting MEAN with Mongo, Express, Angular, and Node*. Shelter Island, NY: Manning Publications.

Holmes, S., and Huber, W. (2019). *Modern Statistics for Modern Biology*. Cambridge: Cambridge University Press, 402.

Hornsby, C., Page, K. M., and Tomlinson, I. P. (2007). What can we learn from the population incidence of cancer? armitage and doll revisited. *Lancet Oncol.* 8, 1030–1038. doi: 10.1016/s1470-2045(07)70343-1

Hu, Q., Sun, W., Wang, C., and Gu, Z. (2016). Recent advances of cocktail chemotherapy by combination drug delivery systems. *Adv. Drug. Deliv. Rev.* 98, 19–34. doi: 10.1016/j.addr.2015.10.022

Lee, E. Y., and Muller, W. J. (2010). Oncogenes and tumor suppressor genes. *Cold Spring Harb. Perspect Biol.* 2:a003236. doi: 10.1101/cshperspect.a003236

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12:323. doi: 10.1186/1471-2105-12-323

Li, X., Cooper, N. G. F., O'Toole, T. E., and Rouchka, E. C. (2020). Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies. *BMC Genomics* 21:75. doi: 10.1186/s12864-020-6502-7

Liu, J., Lichtenberg, T., Hoadley, K., Poisson, L., Lazar, A., Cherniack, A., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.

Lu, D.-Y., Lu, T.-R., Yarla, N. S., Wu, H.-Y., Xu, B., Ding, J., et al. (2017). Drug combination in clinical cancer treatments. *Rev. Recent Clin. Trials* 12, 202–211.

Malvia, S., Bagadi, S. A. R., Pradhan, D., Chintamani, C., Bhatnagar, A., Arora, D., et al. (2019). Study of gene expression profiles of breast cancers in indian women. *Sci. Rep.* 9:10018. doi: 10.1038/s41598-019-46261-1

Masic, I., Miokovic, M., and Muhamedagic, B. (2008). Evidence based medicine – new approaches and challenges. *Acta Inform. Med.* 16, 219–225. doi: 10.5455/aim.2008.16.219-225

McShane, L. M., and Polley, M. Y. (2013). Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical

robustness and clinical utility. *Clin. Trials* 10, 653–665. doi: 10.1177/1740774513499458

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Naito, Y., and Urasaki, T. (2018). Precision medicine in breast cancer. review article. *Chin. Clin. Oncol.* 7:29. doi: 10.21037/cco.2018.06.04

Nikanjam, M., Liu, S., and Kurzrock, R. (2016). Dosing targeted and cytotoxic two-drug combinations: lessons learned from analysis of 24,326 patients reported 2010 through 2013. *Int. J. Cancer* 139, 2135–2141. doi: 10.1002/ijc.30262

Novartis (2020). *The Precision Oncology Annual Trend Report: Perspectives From Oncologists, Pathologists, and Payers. Sixth Edition. 48.* Available online at: https://www.hcp.novartis.com/globalassets/migration-root/hcp/care-management-new/assets/mmo-1224797-the-precision-oncology-annual-trend-report-6th-edition.pdf

Ogino, S., Fuchs, C. S., and Giovannucci, E. (2012). How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. *Expert. Rev. Mol. Diagn.* 12, 621–628. doi: 10.1586/erm.12.46

Oshlack, A., and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* 4:14. doi: 10.1186/1745-6150-4-14

Ozery-Flato, M., Linhart, C., Trakhtenbrot, L., Izraeli, S., and Shamir, R. (2011). Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy. *Genome Biol.* 12:R61. doi: 10.1186/gb-2011-12-6-r61

Preissner, S., Dunkel, M., Hoffmann, M. F., Preissner, S. C., Genov, N., Rong, W. W., et al. (2012). Drug cocktail optimization in chemotherapy of cancer. *PLoS One* 7:e51020. doi: 10.1371/journal.pone.0051020

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 12:480. doi: 10.1186/1471-2105-12-480

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 2004:3.

Teschendorff, A. E., and Severini, S. (2010). Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Syst. Biol.* 4:104. doi: 10.1186/1752-0509-4-104

Tilli, T. M., Carels, N., Tuszynski, J. A., and Pasdar, M. (2016). Validation of a network-based strategy for the optimization of combinatorial target selection in breast cancer therapy: siRNA knockdown of network targets in MDA-MB-231 cells as an in vitro model for inhibition of tumor development. *Oncotarget* 7, 63189–63203. doi: 10.18632/oncotarget.11055

Tom, J. A., Reeder, J., Forrest, W. F., Graham, R. R., Hunkapiller, J., Behrenset, T. W., et al. (2017). Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics* 18:351. doi: 10.1186/s12859-017-1756-z

van Wieringen, W. N., and van der Vaart, A. W. (2011). Statistical analysis of the cancer cell's molecular entropy using high-throughput data. *Bioinformatics* 27, 556–563. doi: 10.1093/bioinformatics/btq704

Verma, M. (2012). Personalized medicine and cancer. *J. Pers. Med.* 2, 1–14. doi: 10.1016/j.pmu.2014.03.007

Vuckovic, N., Vuckovic, B. M., Liu, Y., and Paranjape, K. (2016). *Accelerating Clinical Genomics to Transform Cancer Care*. Santa Clara, CA: Intel.

Welch, B. L. (1949). Further note on Mrs Aspin's tables and on certain approximations to the tabled function. *Biometrika* 36, 293–296.

West, J., Bianconi, G., Severini, S., and Teschendorff, A. E. (2012). Differential network entropy reveals cancer system hallmarks. *Sci. Rep.* 2:802. doi: 10.1038/srep00802

Willems, S. M., Abeln, S., Feenstra, K. A., de Bree, R., van der Poel, E. F., de Jong, R. J. B., et al. (2019). The potential use of big data in oncology. *Oral Oncol.* 98, 8–12. doi: 10.1016/j.oraloncology.2019.09.003

Wilsdon, T., Barron, A., Edwards, G., and Lawlor, R. (2018). *The Benefits of Personalised Medicine to Patients, Society and Healthcare Systems*. Boston, MA: Charles River Associates.

Xia, Y., Fan, C., Hoadley, K. A., Parker, J. S., and Perou, C. M. (2019). Genetic determinants of the molecular portraits of epithelial cancers. *Nat. Commun.* 10:5666. doi: 10.1038/s41467-019-13588-2

Ye, C. J., Regan, S., Liu, G., Alemara, S., and Heng, H. H. (2018). Understanding aneuploidy in cancer through the lens of system inheritance, fuzzy inheritance and emergence of new genome systems. *Mol. Cytogenet.* 11:31. doi: 10.1186/s13039-018-0376-2

Yoder, O. C. (1980). Toxins in pathogenesis. *Annu. Rev. Phythopathol.* 18, 103–129.

Conflict of Interest: The intellectual property of this research is protected by the Brazilian patent number BR1020150308191.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.