AMERICAN UNIVERSITY OF BEIRUT

AN INTEGRATED TCGA PAN-CANCER ANALYSIS ON BIOLOGICAL VARIABILITY ACROSS PATIENTS

by
HEBA TOUFIC FLEIHAN

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Biochemistry and Molecular Genetics
of the Faculty of Medicine
at the American University of Beirut

Beirut, Lebanon
October 2020

AMERICAN UNIVERSITY OF BEIRUT

AN INTEGRATED TCGA PAN-CANCER ANALYSIS ON BIOLOGICAL VARIABILITY ACROSS PATIENTS

by

HEBA TOUFIC FLEIHAN

Approved by: [Signature]

Dr. Pierre Khoueiry
Assistant Professor, Biochemistry and Molecular Genetics
Advisor
[Signature]

Dr. Nadine Darwiche
Professor, Biochemistry and Molecular Genetics
Member of Committee
[Signature]
Nadine Darwiche

Dr. Rami Mahfouz, MD
Professor, Pathology & Laboratory Medicine
Member of Committee
[Signature]

Date of thesis/dissertation defense: 25 November, 2020

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: _____Fleihan___________Heba______________Toufic______________
Last First Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

□ As of the date of submission

□ One year from the date of submission of my thesis.

□ Two years from the date of submission of my thesis.

☒ Three years from the date of submission of my thesis.

___________________________________12/1/2021____________

Signature Date

(This form is signed & dated when submitting the thesis to the University Libraries ScholarWorks)

ACKNOWLEDGMENTS

I would like to express all my respect and gratitude to my advisor Dr. Pierre Khoueiry for providing me the opportunity to complete my master’s project. I truly appreciate his trust in my capacities, and I am thankful for his support and guidance.

I would like to thank my committee members, Dr. Rami Mahfouz and Dr. Nadine Darwiche for reviewing my dissertation and evaluating my work.

I would like to thank my colleagues Mr. Abdallah El Kurdi and Mr. Mohammad Hallal for their encouraging guidance and help.

I would like to acknowledge my parents and my family for their support and for being a constant source of motivation during my studies.

I would like to express my gratitude to my husband for his continuous motivation and endless support.

AN ABSTRACT OF THE THESIS OF

Heba Toufic Fleihan for Master of Science
Major: Biochemistry and Molecular Genetics

Title: An Integrated TCGA Pan-Cancer Analysis on Biological Variability Across Patients

Introduction: Inter-individual cancer variability remains the main challenge for resistance to drug treatment. One of the well-known reasons is the set of genetic alterations and polymorphisms affecting target genes, which cause subsequent changes in gene expression patterns among patients. Whether the observed variability in gene expression affects diagnosis, prognosis, and outcome remain poorly understood. For this, understanding the molecular mechanisms underlying biological variability entails identifying the set of variable and non-variable genes in different cancer types.

Aim: In this study, we hypothesize that biological variability metric will identify key genes that play a role in cancer diagnosis, progression and drug-response.

Methods: Biological variability was calculated on patients' transcriptome data across 33 different cancers retrieved from the TCGA database. Profiling the transcriptome in individuals using RNA-seq technologies has been widely used to obtain mRNA-based molecular markers. Given the robustness of RNA-seq data, we propose a metric that can easily be implemented to detect molecular biomarkers, whether diagnostic, prognostic, or therapeutic using gene expression data...

Results: We derived a list of prognostic and diagnostic markers that were cancer type-specific or common between cancers. We then derived all the list of potential drug-target genes based on their biological variability score and oncogenic properties

Conclusion: Not only is biological variability an important measure to identify variable and non-variable genes in each cancer, but it is key to identify cancer-specific molecular markers, predict reliable drug-target genes and identify genes related to cancer progression and development. Gaining a deeper understanding of genes expression variability in cancer will broaden our knowledge on genes related to resistance early cancer detection, outcome, and the development of personalized treatment.

CONTENTS

ILLUSTRATIONS

Figure

ABBREVIATIONS

ACC	Adrenocortical carcinoma
ACTB	Actin Beta
ADAM7	ADAM Metallopeptidase Domain 7
APO	Apolipoprotein A
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CALR	Calreticulin
CEACAM5	CEA Cell Adhesion Molecule 5
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHGA	Chromogranin A
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
CV	Coefficient of Variance
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
EIF4A2	Eukaryotic initiation factor 4A2
ESCA	Esophageal carcinoma
FAM72B	Family With Sequence Similarity 72 Member B
FFPE	Formalin-Fixed Paraffin-Embedded
FTL	Ferritin Light Chain
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GDSC	The Genomics of Drug Sensitivity in Cancer
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma

IGF2	Insulin Growth Factor 2
KICH	Kidney Chromophobe Carcinoma
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
KLK3	Kallikrein Related Peptidase 3
KRT6A	Keratin 6A
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MCL1	MCL1 Apoptosis Regulator, BCL2 Family Member
MESO	Mesothelioma
ORM1	Orosomucoid 1
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PIP	Prolactin-induced protein
PCPG	Pheochromocytoma and Paraganglioma
PRAC1	Prostate Cancer Susceptibility Candidate Protein 1
PRAD	Prostate adenocarcinoma
PSA	Prostate Specific Antigen
READ	Rectum adenocarcinoma
REG1A	Regenerating Family Member 1 Alpha
REG3A	Regenerating Family Member 3 Alpha
Rlog	regularized log transformation
RPL7A	60S ribosomal protein L7a

RPL8	60S ribosomal protein L8
SARC	Sarcoma
SERPINB3	Serpin Family B Member 3
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCA	The Cancer Genome Atlas
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma
THYM	Thymoma
TMSB10	Thymosin Beta 10
TPM	Transcript Per Million
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UPK	uroplakin 2
UVM	Uveal Melanoma
VST	variance stabilizing transformation

CHAPTER I

INTRODUCTION

With the race towards precision medicine, the past decade has witnessed a rapid acceleration in our understanding of the genetic basis of cancer growth and development. This was followed by redefining the drug-targeting approaches and moving towards personalized treatment [1]. However, the latter was faced with several challenges resulting from the inter-individual heterogeneity. This variability is the result of the inherently unstable nature of cancer, characterized by a set of genetic alterations in gene expression patterns that stems from various molecular aberrations that occur in the disease course [2].

A key factor for determining inter-individual heterogeneity is studying the variability of expression of key genes that play a role in cancer progression and prognosis. For example, transcriptome data of patients with Adenoid Cystic Carcinoma revealed that 20% of the patients had a poor survival rate, due to a set of genes that resembled embryonic stem cells [3]. This variability of expression stems from the stochastic nature of genes, which is thought to be the consequence of several epigenetic and regulatory factors in the genome [4]. Moreover, a multivariate analysis used on transcriptome data of Glioblastoma patients identified a set of prognostic markers that played a significant role in disease progression and tumorigenesis [5].

The variability of drug effectiveness and response is due to the variability of expression of cancer biomarkers and drug-targeted genes [6]. For example, molecular profiling of prostate cancer biopsies from treated patients identified 7 gene signature biomarkers, including a gene that exhibits a chemo-resistant property (ORM1) and two cell-cycle related genes that serve as prognostic markers (ADAM7 and FAM72B) [7]. Another study by Simonovsky et al identified genetic polymorphism in drug-targeted genes and discovered that drugs targeting variable genes tend to be ineffective in the population [6].

With the rapid acceleration of high through-put technologies, several projects have emerged like The Cancer Genome Atlas (TCGA) which generated a large amount of transcriptomic and epigenetic data [8], and hence created an opportunity for bioinformaticians to mine this data and turn it into a valuable recourse. It also allowed researchers to explore molecular profiling data through a comprehensive and integrative pan-cancer analysis. Zhang et al carried a systematic and pan-cancer epigenetic analysis on 11 cancer types and identified co-methylation clusters in 11 cancer types [9]. Another study by Hoadley et al conducted a comprehensive pan-cancer analysis and identified a common subtype characterized by TP53 alterations and deregulation of immune gene signatures [10]. More recently, Cao et al investigated common and specific cancer signatures related to immune response, cell cycle and angiogenesis through conducting a comparative pan-cancer analysis [11]. However, understanding how gene expression variability correlates to variation drug response and specific cancer signatures is still a challenge.

RNA-seq data has been widely used in countless areas of cancer research to profile the composition of the entire transcriptome, including coding and non-coding RNAs [12, 13]. It has helped identify a wide range of therapeutic gene targets, biomarkers, and gene expression patterns among patients. This means that RNA sequencing can reveal a wide range of functional and structural changes experienced by genes rather than just revealing specific mutations. For example, Biton et al used bladder cancer transcriptome data to identify components related to tumor environment and tumorigenesis [14]. However, to our knowledge there is no large-scale pan-cancer study that explores cancer diagnostic and prognostic biomarkers using gene expression variability. Here, we hypothesize that performing a systematic pan-cancer analysis on biologically variable and non-variable genes will help us in discovering genes that play role in cancer diagnosis and prognosis.

In this study, we used RNA-seq of 11,000+ patients from 33 different cancer types from the TCGA database to analyze gene expression variability [8]. We derived a metric to extract biological variability from the observed total variability in gene expression between patients. This metric allowed us to identify cancer types sharing similar patterns in gene expression variability. Additionally, we detected association between biological variability and molecular biomarkers.

CHAPTER II

THESIS OBJECTIVES AND AIMS

Aims:

1. Derive a metric to call biologically variable genes based on RNA expression among patients of the same cancer type.
2. Conduct a pan-cancer analysis on transcriptome data from the TCGA database to identify biologically variable and non-variable genes in different cancer types.
3. Extract cancer-specific RNA-based diagnostic and prognostic markers from biologically variable genes.
4. Identify potential drug-target genes based on biological variability score and oncogenic property.

CHAPTER III

MATERIALS AND METHODS

A. Data Processing

Data were downloaded from the TCGA database using custom R/Bioconductor scripts and duplicated samples were removed. Some of the primary tumor samples included both Formalin-Fixed Paraffin-Embedded (FFPE) and primary solid tumor samples from the same patient. Given that FFPE-derived RNA is highly degraded which in turn impacts its efficacy as a reliable source for biological analysis [15], we chose to proceed with fresh primary solid tumor samples for solid tumors and peripheral blood samples for liquid tumors. For this, we used the TCGABioLinks R package (https://bioconductor.org/packages/TCGAbiolinks/) to get information on the type of biospecimen, whether it is an FFPE or not. Next, we removed duplicated samples using a source code available on Github (ShixiangWang/ tcga_replicateFilter.R) [16]

B. Data Normalization and Transformation

We performed three normalization techniques in order to choose the best one that fits our experimental design. One is Transcript Per Million (TPM) normalization, a within-sample normalization which uses a biological approach to normalize data by taking into account the transcript size [17]. TPM is calculated using the following formula:

$$RPK\ ij = \frac{\mathrm{Rij}}{Li}$$

$$\mathrm{TPM\ ij} = \mathrm{RPK\ ij}\ \div \frac{\sum RPK\ ij}{1{,}000{,}000}$$

where Rij is Read Count of gene ***i*** in patient ***j*** and L***i*** is the length of gene ***i*** in kilobase pair. [18]. The second method is log2 (TPM +1), a highly used normalization technique in several areas of research. The third and fourth methods are vst (variance stabilizing transformation) and rlog (regularized log transformation), two transformation techniques acquired from the DESeq package [19]. Rlog performs the same normalization as vst but it takes a longer time to compute when analyzing data greater than 100 samples; therefore, we proceeded with vst transformation since most of the cancer types had more than 100 samples.

In calculating gene expression variability, we need to find the variable and non-variable genes regardless of gene expression level. For this we used the SCnorm package [20] which computes gene count-depth relationship and determines which of the techniques mentioned above is unbiased to any gene expression difference. When using SCnorm, genes that had an expression of zero in more than 20% samples were removed. The methods were then compared in biological context to check which of the methods help control false discoveries.

C. Evaluating different normalization and transformation techniques

For the normalization, we adopted one of the most used techniques, called Transcripts Per Million or TPM, that scale all libraries to 1,000,000 reads and takes into consideration the gene's length. We then used the SCnorm package [20] to evaluate the effectiveness of two widely used transformation techniques, the log2 and vst (Varying Stabilizing Transformation). In each case, the slope of gene ***i*** is calculated by finding the ratio of count data ***y*** and the sequencing depth ***x*** between patients.

$$slope(i) = \frac{\mathrm{yb} - \mathrm{ya}}{\mathrm{xb} - \mathrm{xb}}$$

D. Filtering Following Biological Variability Calculation

Once we collected the biologically variable genes, we filtered out the non-variable genes (Biological variability < 0) that had a mean expression of at most 0.5 following normalization. Then, we removed the genes that had a length of less than 200 kb to filter out all the non-coding RNA and we excluded the X &Y genes in our analysis so that they won't interfere in the biological variability results.

Genes biotype was determined using the Ensembl Biomart [21], which stratified the genes into protein-coding, non-coding and pseudogenes. The latter were filtered out during the analysis. Genes that had a biological variability less than or equal to zero were considered non-variable and genes with a biological variability greater than 0 were variable.

E. Cancer type-specific diagnostic markers

Given a cancer type, we focused on the least variable genes to extract the diagnostic markers that are gene-cancer specific by comparing their variability score in the respective cancer to their score in other cancer types. Ideally, biomarkers should be cancer-specific and possess similar gene expression patterns between cancer patients [22]. Cancer-specific diagnostic markers were extracted by exploring the protein-coding genes that possessed a biological variability score less than -5 in one cancer and variable in other cancers. We validated our results by comparing them to Clinical Interpretation of Variants in Cancer (CIViC) Database (http://bionlp.bcgsc.ca/civicmine/). This database uses text-mining approaches to extract all the clinically relevant biomarkers from published articles [23].

F. Extracting Prognostic Markers

Prognostic markers are more prone to genetic variability and may predispose to treatment response. This is because prognostic markers reflect the tumor stage and give us insights to possible remissions and recurrence of cancers [24]. Prognostic markers are identified by looking at the alterations in gene expression patterns of protein-coding genes related to cancer progression and proliferation [24]. Therefore, to extract the prognostic markers, we looked for the variable genes with scores greater than 4 that are either cancer-specific or common in no more than 6 cancers. We also validated our results using the CIViC Database [23].

G. Extracting drug-target genes

First, we got all FDA-approved drugs with their drug-target genes from the Genomics of Drug Sensitivity in Cancer (GDSC) database (https://www.cancerrxgene.org/) [25]. Then, we checked the oncogenic property (oncogene, TSG, oncogene/TSG) of the drug target genes using the OncoKB database (https://www.oncokb.org/cancerGenes) [26].

To extract the potential drug-target genes from our data, we annotated the genes based on their oncogenic property; then we picked the oncogenes that have a low variability score in several cancers.

CHAPTER IV

RESULTS

A. Compiling RNA-seq data for 33 cancer types from the TCGA database

Transcriptome data profiles of human tumor samples were obtained from The Cancer Genome Atlas (TCGA) database GDC portal [8]. In July 2019, there were 11,315 samples, spanning 33 cancer types. Primary tumor and matched normal tissue samples were selected in each cohort as represented in figure 1. Only 21 cancer types have normal samples, of which only 16 cancers have more than 10 matched normal tissue samples. To derive a metric for calling biologically variable genes, we used the primary tumor samples owing their large number of samples in different cancer types.

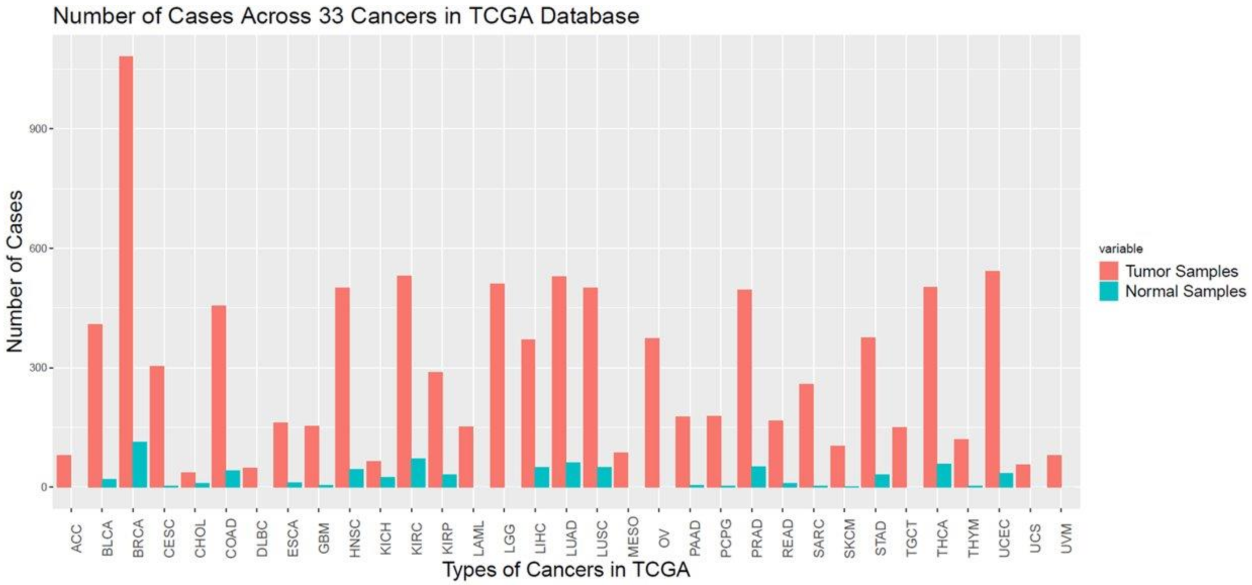

Figure 1. Number of tumor and normal samples in each of the TCGA cancers.

In this study we processed and downloaded harmonized raw count RNA-seq data (data aligned to the most updated human gene assembly, hg38) and processed it using R/Bioconductor programming language with a total of 56,734 coding and non-coding genes in each cancer. Raw data was filtered for FFPE samples (Materials and Methods).

B. Data Normalization and Transformation

Raw transcriptome data may vary due to some uncontrolled experimental conditions, also known as batch effects [27][28]. Consequently, it must be normalized and transformed prior to any analysis so that genes belong to different samples and expression levels can be compared. Essentially, normalization corrects for differences in sequencing depth between RNA-seq libraries/samples and for gene lengths while transformation alleviates differences between gene groups with varying expression levels. Therefore, choosing the best normalization and transformation methods help reducing systematic-derived variability, prevent biased results in our analysis, and make transcriptome data comparable across samples.

To this end, we first tested several combinations of normalization and transformation techniques and evaluated them by estimating the gene count-depth relationship on 10 equally-sized expression groups ranging from low to high [20] (Material and Methods, Fig. 2 and 3). The test on different expression groups will allow us to identify the method that has no bias to gene expression levels.

The method is considered to be effective if the slope of the count-depth for all expression levels is flat, or near 0 in mathematical terms. Figure 2 shows the slopes of 3 different expression groups (low, medium and high) in a selection of 6 different cancers for clarity, chosen randomly out of the 33 cancers: *BRCA* (Breast invasive carcinoma), *ACC* (Adrenocortical carcinoma), *KICH* (Kidney Chromophobe Carcinoma), *BLCA* (Bladder Urothelial Carcinoma) , *COAD* (Colon Adenocarcinoma) , and *LGG* (Brain Lower Grade Glioma). The 3 expression groups fell respectively in the 10-20th, 40-50th and 50-70th quantiles. Prior to normalization, the slopes of the raw count data in all of the expression groups were greater than 0 (non-flat; Fig 2A). In contrast, both log2 (TPM+1) and vst methods showed slopes closer or equal to zero in all of the expression groups (Fig 2C-D), as opposed to TPM alone (Fig. 2B).

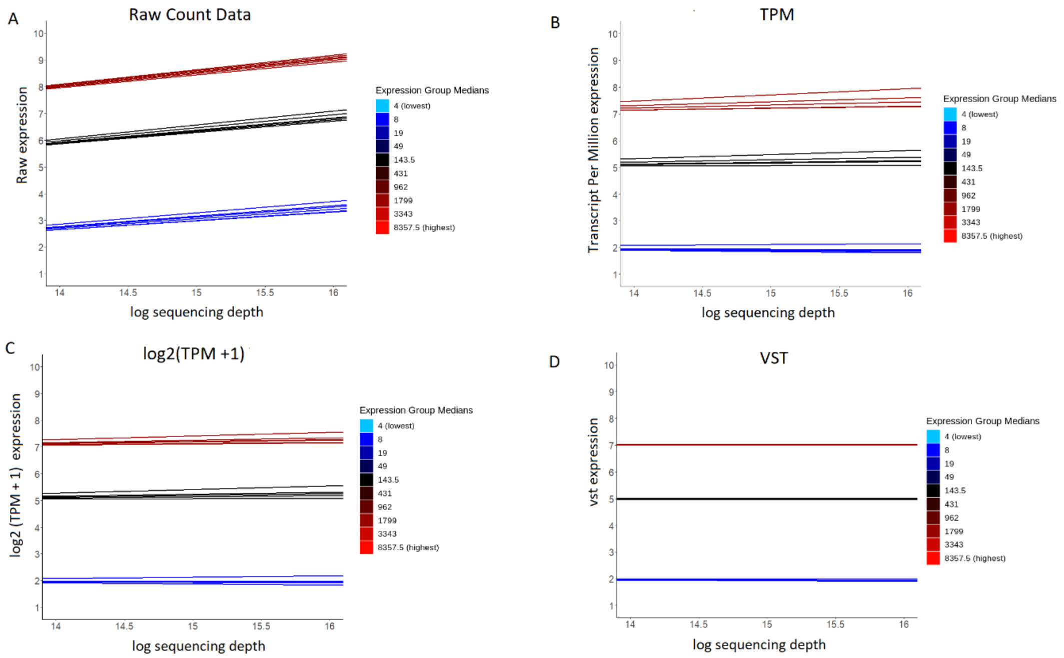

Figure 2. Comparison of the Count Sequencing Depth Slopes between different normalization techniques. Slopes for log count data and sequencing depth for 3 different expression groups, low (blue), moderate (black) and high (red), determined by their median expression in each quantile in 6 different cancers. Slopes of the genes that fall in each of the 3 quantiles were averaged.

To further confirm our findings, we calculated the densities of the slopes within each of the 10 equally sized expression groups for the different normalization techniques on one cancer type, the KICH cancer dataset. Again, both log2 (TPM +1) and vst (Fig. 3C and D respectively) were successful in reducing the variability between different expression groups as compared to raw counts or TPM alone (Fig. 3 A and B, respectively).

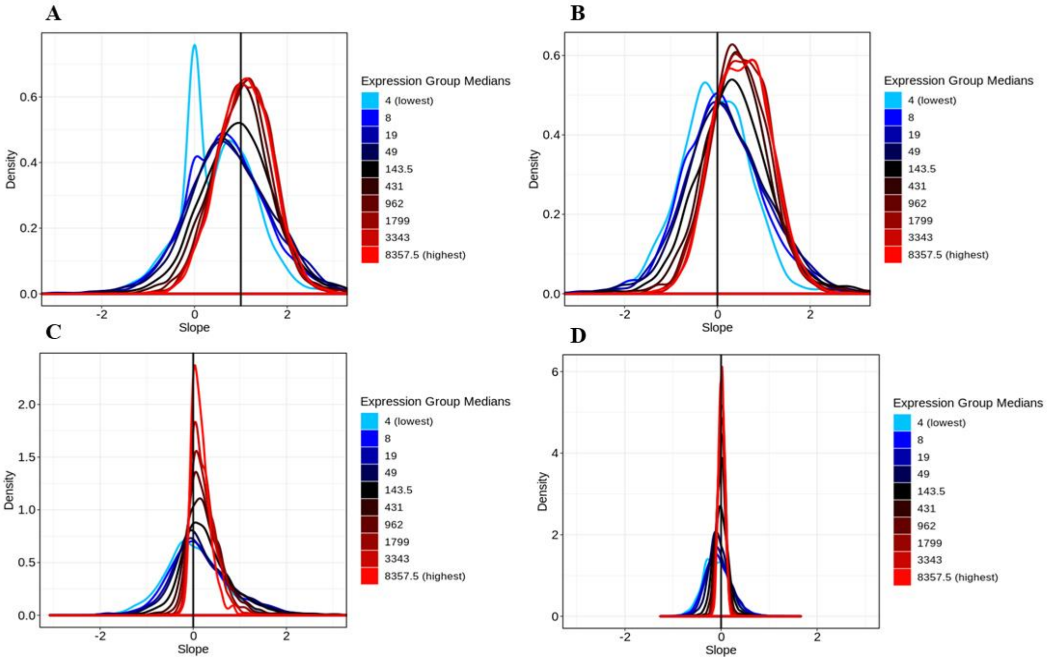

Figure 3. Comparison of the densities of slopes of 10-equal expression group for different normalization techniques. Density plots of the count depth relationship of 10 equally sized gene groups (A) Unnormalized raw data from the KICH dataset (B) Data normalized using TPM (C) Data normalized using log2(TPM+1) (D) Data normalized using vst

In conclusion, both log2(TPM+1) and vst methods were effective in reducing technical variabilities between genes and thereby eliminating batch effect. However, since vst is slow to calculate for large number of genes and is considered to be conservative, we adopted the log2(TPM +1) as a method for normalizing and transforming TCGA gene expression data.

C. Estimating Biological variability from normalized and transformed RNA-seq data

Inter-tumor heterogeneity remains a major challenge in cancer research. This is due to due to genetic polymorphism in key genes that play a role in cancer progression and proliferation; hence affecting response to treatment. In order to identify these genes, we measured the biological variability of transcriptome data across all samples/patients of a cancer for each of the 33 cancers from the TCGA database. While CV (coefficient of variance) and total variance were previously used to calculate gene expression variability between patients, the former has shown to be biased towards low expressed genes and the later towards highly expressed genes [29][30].

Total variability in gene expression can be due to three factors: biological variability (1), technical variability (2), and the variability that arise from shot or counting noise (or sampling during the experimental procedure) (3) (Fig 4).

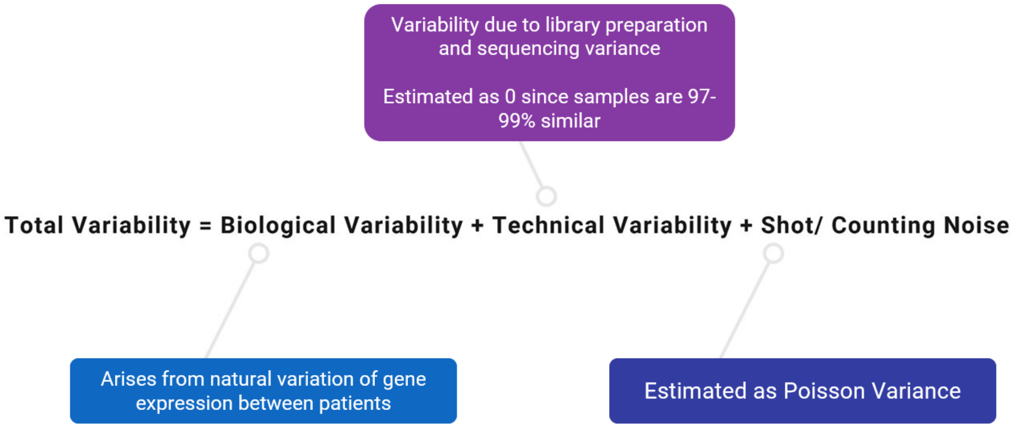
Variability due to library preparation and sequencing variance
Estimated as 0 since samples are 97-99% similar
Total Variability = Biological Variability + Technical Variability + Shot/ Counting Noise
Arises from natural variation of gene expression between patients
Estimated as Poisson Variance

Figure 4. Variability in gene expression arises from 3 forms of variability (biological variability, technical variability and shot noise)

Total variability is mathematically defined as the average of the squared differences from the Mean. For a gene x, total variability will correspond to:

$$S^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (x_i - \overline{x})^{\,2}}{n-1}$$

Biological variability is the variability we intend to measure; it stems from the natural variation of gene expression patterns between patients.

Technical variability is the variability that arises from sequencing variance during library preparation steps; it represents the random variability between technical replicates associated with experimental procedures like amplification, reverse transcription, and RNA extraction. However, RNA-seq experiments proved to have excellent technical reproducibility, and it was shown that technical replicates are 97- 99% similar [31]. Taking this into account, we assume the technical variability to be equal to zero.

Last, **shot or sequencing noise variability** emerges from the uncertainty in measuring count data; it’s the unavoidable noise that arises from RNA-seq experiments and recurs even if everything in the experimental procedure is the same [32]. For example, even if the transcripts' and aliquots' concentrations are equal in the flow cell lanes, the count data will still vary, and this variability is known as shot noise [32]. **The shot noise follows a Poisson distribution** since, unlike microarrays that measure continuous data, RNA-seq measures count data, and therefore can't be assumed to follow a normal distribution (Fig.

5). Poisson distribution considers each individual piece of mRNA to be randomly drawn from a pool of mRNA.

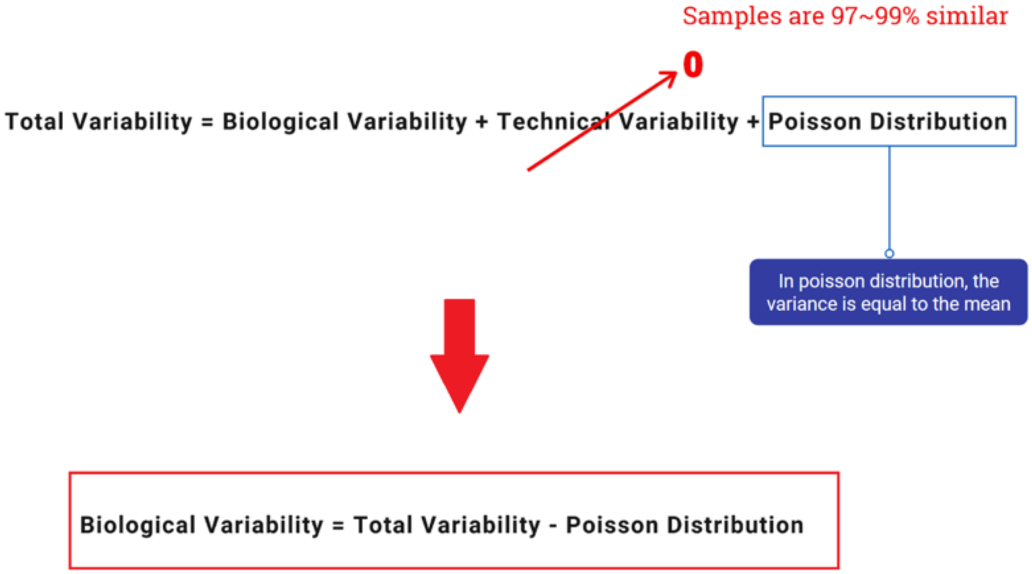
Samples are 97~99% similar
0
Total Variability = Biological Variability + Technical Variability + Poisson Distribution
In poisson distribution, the variance is equal to the mean
Biological Variability = Total Variability - Poisson Distribution

Figure 5. Estimating biological variability. Technical variability is estimated as 0 since technical replicates are 97-99% similar. Shot Noise follows a Poisson distribution in which the variance is equal to the mean.

As such, in mathematical terms, the biological variability of a gene expression can be obtained using the formula where μX_i corresponds to the Poisson distribution variance:

$$Var\ Xi = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^{2}}{n-1} - \mu Xi$$

Using the above formula, we estimated the biological variability of all genes in each of the 33 cancer types. We then classified the genes into two categories based on their biological variability measure: 1) “**Variable**” for genes with variability > 0; 2) “**Non-Variable**” for genes with variability <= 0. We suggest that genes classified in each category pose certain biological and functional roles in cancer development and detection.

D. Biological variability is cancer specific

To gain an idea about the extent of biologically variable genes in different cancers, we compared the range of variability for each cancer in the 33 different cancer types. As shown in Fig. 6. Glioblastoma (GBM) and Skin Cutaneous Melanoma (SKCM) were the cancers with the lowest and highest range, respectively. Interestingly, the high number of variable genes observed in SKCM may explain to the heterogeneity and plasticity of the tumor [33].

We notice that several cancers share the same lowest non-variable gene. For example, FTL gene encodes Ferretin light chain, a protein that is important for iron homeostasis and TMSB10 (Thymosin Beta 10) plays an important role in cytoskeleton organization. These genes along with other housekeeping genes like GAPDH, ACTB and ribosomal proteins like RPL8 (60S ribosomal protein L8) and RPL7A (60S ribosomal protein L7a) are genes with low variability scores most likely related to their basic cellular functions and thus justifies their use as reference genes since they are stably expressed and consistent across all cancer types [34].

On the other hand, each cancer has its unique highly variable gene. For example, PRAC1 is the gene with the highest variability in Colon Adenocarcinoma. Interestingly, PRAC1 is known for being deferentially expressed between right sided and left-sided colon cancer [35]. Another example is the orosomucoid 1 gene (ORM1) in Prostate Adenocarcinoma, a highly biologically variable gene promoting prostate cancer metastasis through its involvement in cancer metabolism and immune response activity [36].

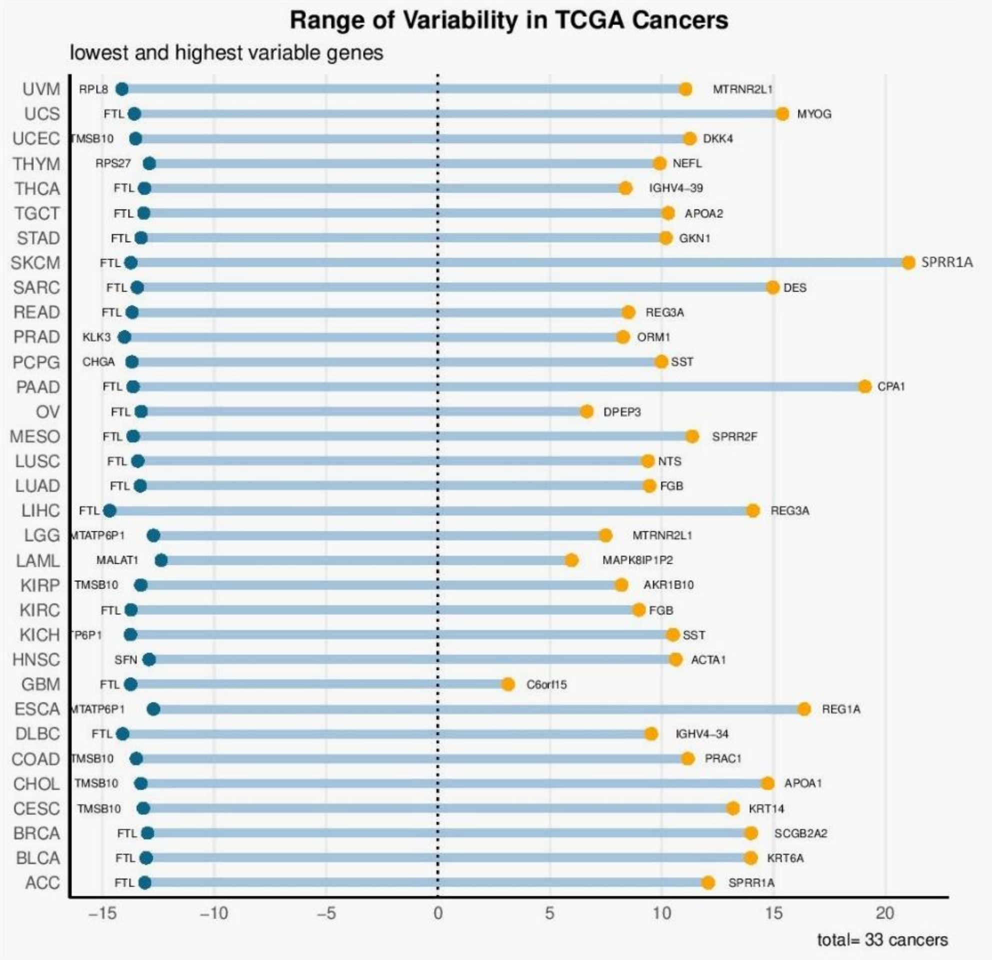

Figure 6. Range of biological variability between cancers. Plot showing the least variable gene in gene cancer and the highest variable gene with their respective biological variability score in each cancer. The blue dots represent the least variable gene and the orange dot represents the highest variable gene in each cancer. The dashed line is the cutoff between variable and non-variable genes (Biological Variability =0).

E. Pan-Cancer Analysis provides further context into cancer-specific biological variability

Systematic pan cancer analysis using multi-omics data has been the new paradigm to understand individual cancers and to extract information based on transcriptome profiles beyond tissue-of-origin cancer classification [37]. Several pan-cancer studies have identified gene networks signatures of prognostic and diagnostic properties [38][39].

We started by performing cross-tumor clustering to examine whether biological variability is tumor specific and to extract common or cancer specific biological variability signatures. For this we collected the list of biologically variable **protein-coding genes** for each of the 33 cancers and clustered the cancers based on their gene expression variability (Fig. 7). Interestingly, cancers with the same histological origin exhibited similar gene signature patterns and were clustered together (Fig. 7A). For instance, Kidney Chromophobe Carcinoma (KICH), Kidney renal clear cell carcinoma (KIRC) and kidney renal papillary cell carcinoma (KIRP) were clustered together indicating that they share similar biological variability profiles. Similarly, for Colon adenocarcinoma (COAD) with Rectum adenocarcinoma (READ) and Brain Lower Grade Glioma (LGG) with Glioblastoma (GBM); However, the last 2 cancers possess a unique gene variability pattern, different from all the other solid tumors, as they formed an outer group to all other cancer types. To further explore this similarity, we calculated the correlation in biological variability for all pairs of cancers and plotted it (Fig. 7B). The Correlation analysis of gene expression variability confirmed the similarity observed above and, additionally, showed that Acute Myeloid Leukemia (LAML) possesses a biological variability pattern that does not match with any of the TCGA cancers (Fig 7B). Given that expression variability was tumor-type specific, this suggests the existence of potential mRNA-based biomarkers that could be detected using biological variability.

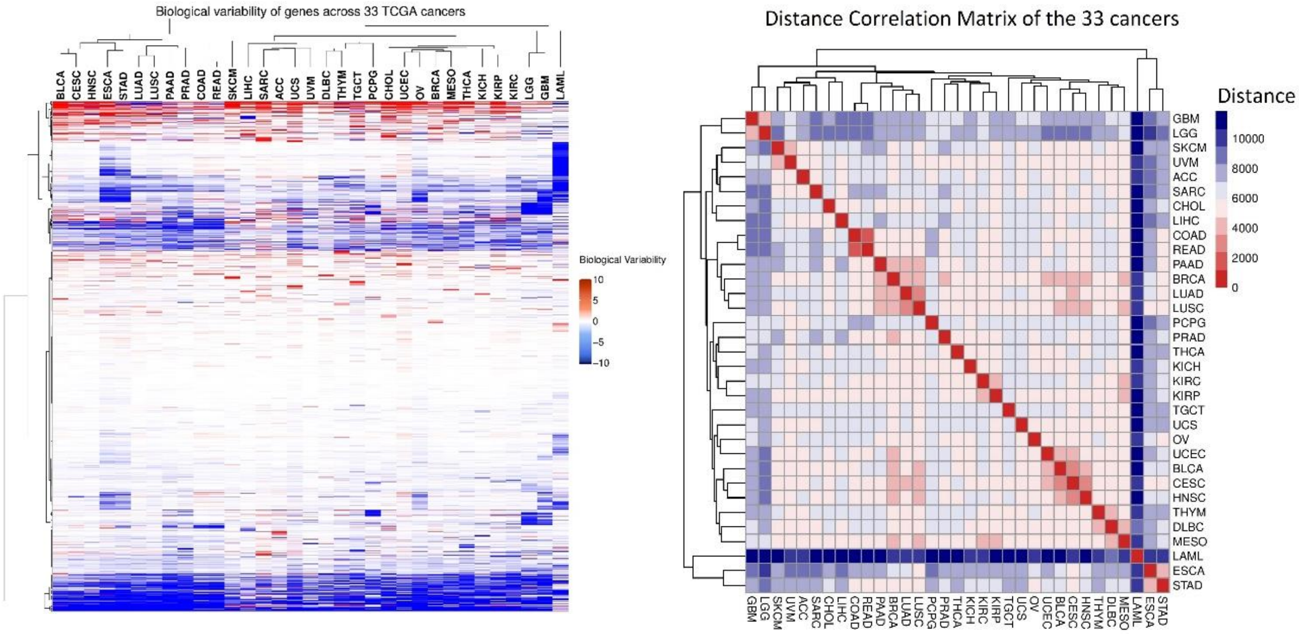

Figure 7. Correlation between the 33 cancers based on biological variability measure. **(A)** Hierarchal clustering of the TCGA cancers based on biological variability score. Rows and columns represent genes and cancer types, respectively. **(B)** Hierarchical distance clustering matrix analysis of the TCGA cancers obtained by using the Manhattan distance method. This shows how similar or dissimilar are cancer types in terms of biological variability in genes. Red highlights pair of cancers with high similarity in their biological variability. LAML shows increased distance (blue squares, high dissimilarity) with all other cancer types highlighting its uniqueness.

Last, we checked whether variable or non-variable genes are more or less common between cancer types. For this, we calculated the frequency of shared high or low variable genes among the 33 cancer types (Fig. 8). Our data showed that the majority of genes with high biological variability are unique to one cancer and that several of the least variable genes were found common between cancers (Fig. 8). For instance, approximately 285 high variable genes (red) are shared among 3 cancer types as opposed to only 39 low variable genes.

Again, this shows that genes biological variability is highly cancer-specific and its use will allow us to unravel cancer specific biomarkers including prognosis and diagnosis genes.

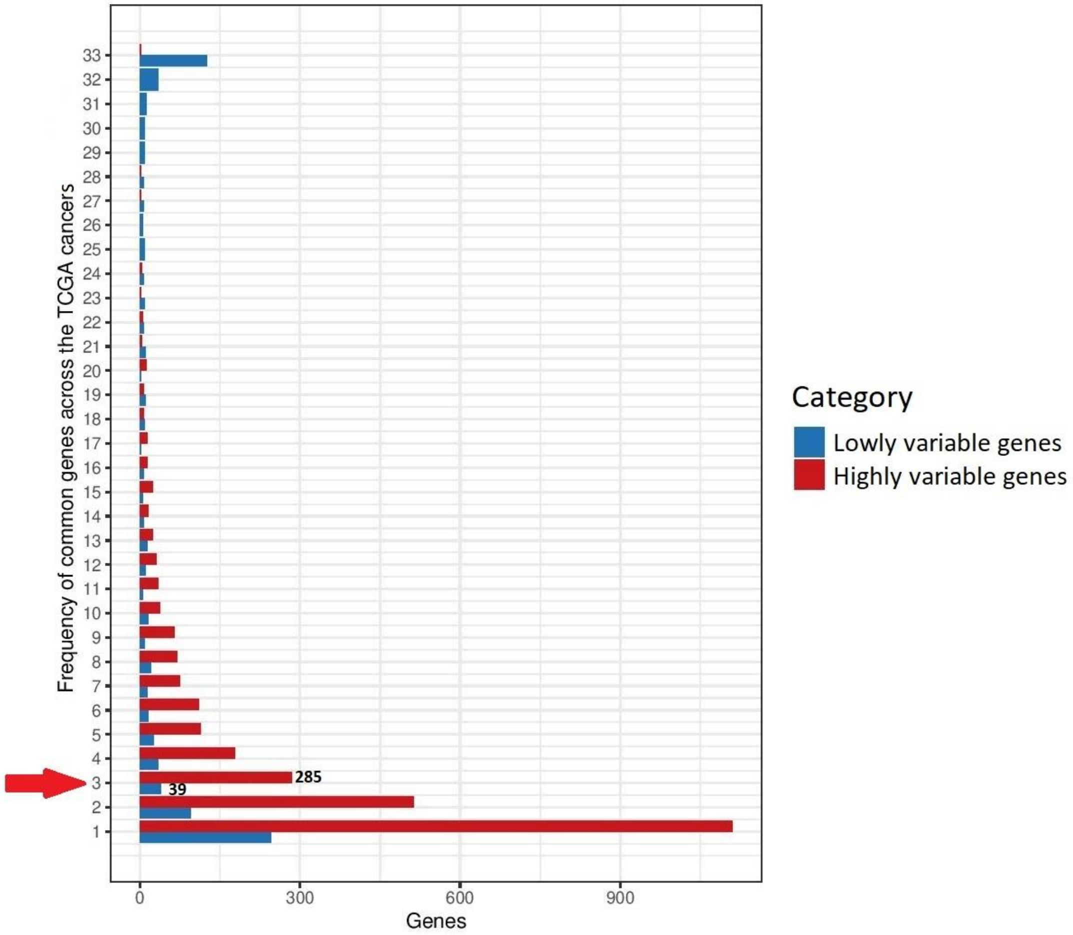

Figure 8. Frequency of common cross-variable genes. The y-axis represents the frequency of occurrence of low variable genes (blue) and high variable genes (red) across all the 33 cancers. The x-axis represents the count of genes. The number of common variable genes increases as the frequency of their occurrence as variable between cancers decreases. As for the non-variable genes, the number of common genes seems to interplay with highest counts at frequency = 1 (only non-variable in one cancer) or 33 (non-variable in all of the 33 cancers). For example, 285 genes appear as variable in 3 cancers while only 39 genes appear as non-variable.

F. Genes encoding diagnostic molecular markers are non-variable.

Our previous data highlighted the cancer-specific nature of biological variability. We thus asked whether cancer-specific non-variable genes are enriched for diagnostic markers. Naturally, for a gene to be classified as diagnostic marker, it has to show low biological variability among patients of the same cancer type. Therefore, we filtered for genes with a low biological variability score in a specific cancer (biological variability < -5) which resulted in a total of 131 genes. We then represented the data in a heatmap showing diagnostic molecular markers in rows and cancer types in columns (Fig. 9). This approach allowed the identification of genes that lack biological variability in one cancer and for which the gene is a characterized diagnostic marker.

For instance, KLK3 is a gene that encodes a prostate-specific antigen (PSA), the biomarker for prostate adenocarcinoma [40] and showed to be exclusively non-variable in prostate adenocarcinoma (PRAD) with a biological variability score of -14.5, while it was variable in the remaining 32 cancers (Fig. 9).

In addition to KLK3, we derived several biomarkers for other cancers, among them were KRT6A, a biomarker for Head and Neck Carcinoma; CHGA, a biomarker for Pheochromocytoma and Paraganglioma; APOA2 and APOA1, biomarkers for Liver hepatocellular carcinoma and many others... (Fig.9, Supplementary Table 1). While some diagnostic markers were specific to one cancer, like the one mentioned above, others were found in more than one cancer, specifically those with the same histological origin. For example, READ and COAD share the same diagnostic markers, including an important and well-studied gene: CEACAM5 (Fig. 9). This gene encodes carcinoembryonic antigen

(CEA), a classical diagnostic marker for colorectal cancer [41]. The same goes for LGG and GBM that share several low variability genes. However, we also notice in this case the presence of genes with low variability in LGG compared to GBM (i.e. sox1) which suggest that our metric allows the identification of diagnostic markers able to differentiate between very similar cancer types.

Last, we crossed compared the list of 131 genes identified here with the list of diagnostic markers from the CIViC database (Material and Methods) and identified 104 common hits. This shows that the biological variability metric proposed here is able to identify known and *de-novo* diagnostic markers.

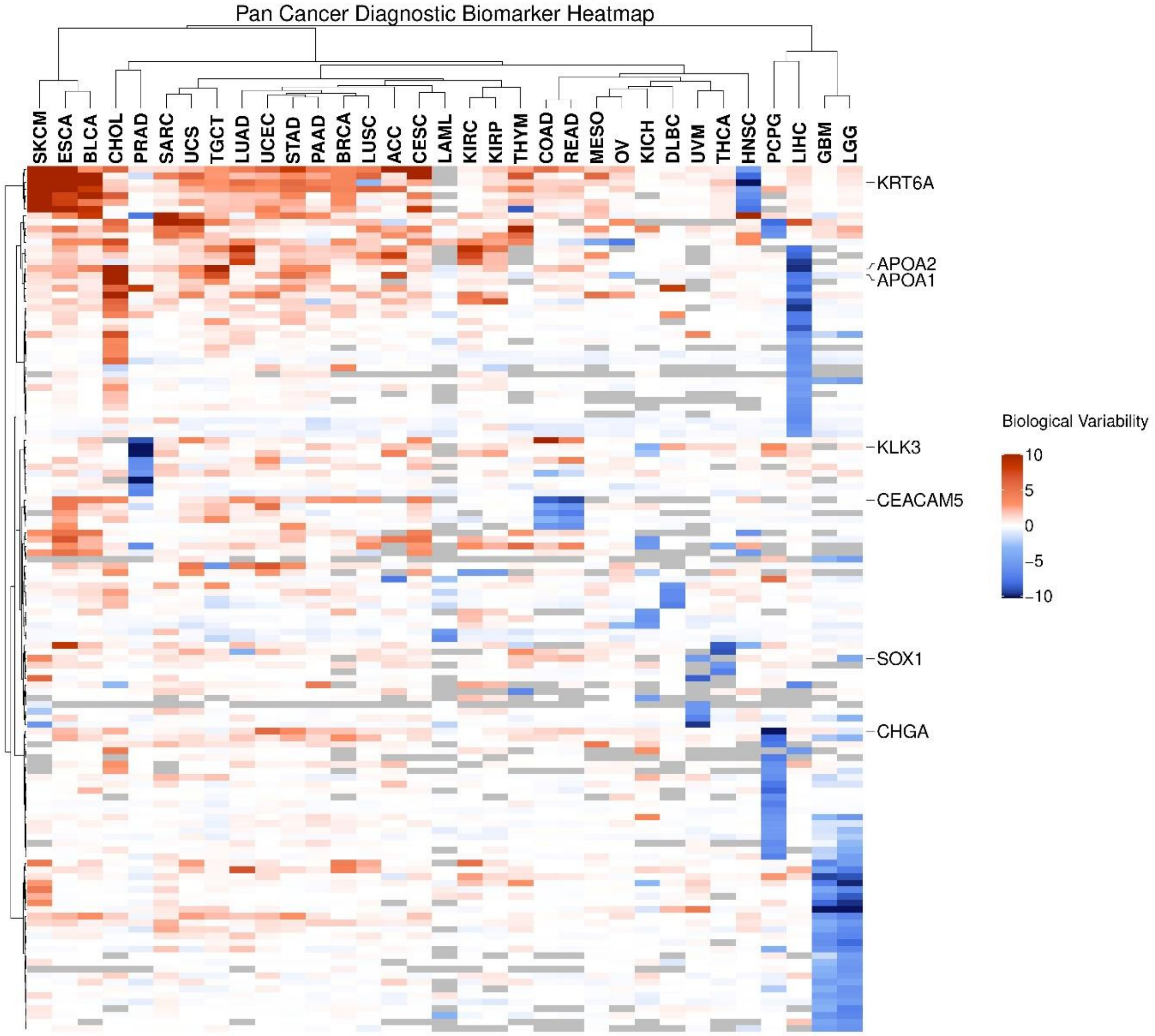

Figure 9. Detecting diagnostic molecular markers. Heatmap showing genes (rows) that are exclusively non-variable (score < -5) in a specific cancer type (columns). Blue squares indicate that the gene is not variable at the specific cancer type. Note that a non-variable gene in a specific cancer type may exhibit high variability in other cancer types which justify the scale range from -10 to +10.

G. Genes encoding prognostic molecular markers are variable.

Since clinical variables such as disease stages are not enough to predict cancer outcomes, even in patients that possess similar clinicopathological characteristics, identifying cancer-specific prognostic markers as well as tumor progression markers based on gene expression variability is necessary to provide evidence-based treatment decisions and improve clinical outcomes [42]. A reliable prognostic marker can provide cancer-stage information and measure the risk of disease progression [43]. Prognostic markers are characterized by a set of epigenetic and genetic alterations of genes corresponding to cancer development and proliferation that influence the disease outcome [24]. The expression levels of these markers progressively change with the progression of cancer [44]. Some prognostic markers may be cancer-specific like UPK 2 (uroplakin 2) in bladder cancer [45] while others may serve as common prognostic markers like IGF2 (Insulin growth factor 2) [46].

In order to identify genes with a putative prognostic and predictive outcome, we extracted highly variable genes with scores greater than 4 and we narrowed down our analysis to those that are only common in less than 6 cancers (Fig. 10, Supplementary Table 1). This resulted in a list of 473 putative prognostic markers identified based on biological variability.

For instance, REG3A and REG1A are two well studied prognostic markers for colorectal cancer that showed increased biological variability in colorectal, rectal and pancreatic adenocarcinomas. Overexpression of these genes activates AKT and ERK 1/2 pathways and promotes tumor proliferation [47]. Another example is PIP (Prolactin-induced protein), a highly biologically variable gene in breast cancer whose targets were shown recently to be related to poor response to chemotherapy [48]. Same for IGF2 (insulin-like growth factor 2), a highly variable oncogene in liver cancer and whose overexpression accelerates liver tumor formation [49].

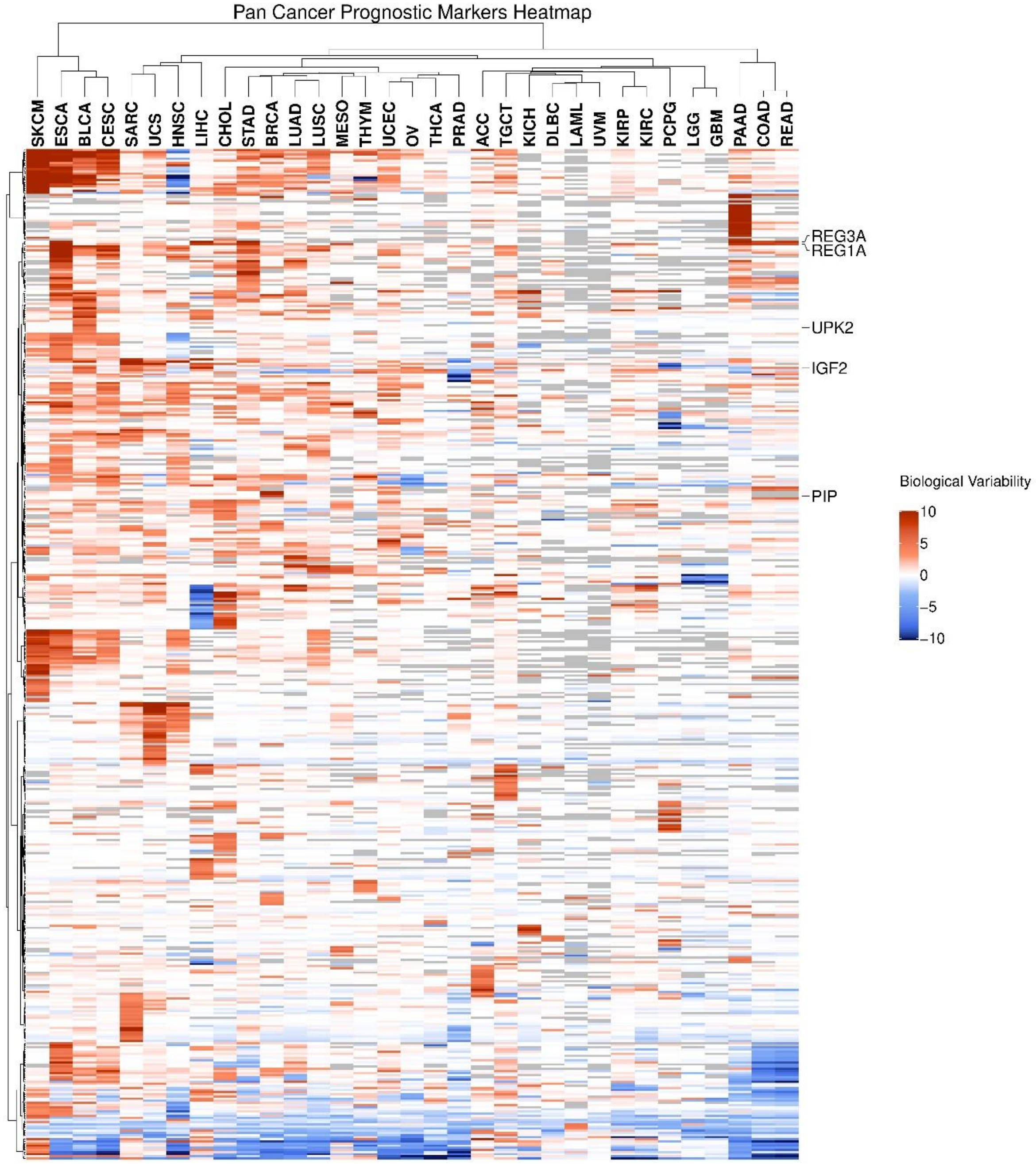

Figure 10. Detecting prognostic molecular markers based on biological variability of gene expression. Genes (row) with a biological variability scores greater than 4 where plotted for each of the 33 cancer types (columns). Note that a gene with a high variability score in one

cancer can exhibit a low variability score in other cancers hence the scale range from -10 to +10. Red depicts genes with high variability scores and blue those with a low variability score.

Last, similar to the analysis in the previous section, we crossed compared the list of 473 genes identified with the list of prognostic markers from the CIViC database (Material and Methods) and identified 99 common hits.

In summary, both diagnostic and prognostic analysis above show that using biological variability helps us identify novel RNA-based diagnostic and prognostic markers that can assist in tumor early detection and outcome.

H. Drug Efficacy Based on Gene Variability

Genetic polymorphism, mutations, and copy number alterations are all factors that affect gene expression levels which in turn cause a heterogeneous response to cancer drug therapy [50]. Drug efficacy has been directly linked to gene expression variability; In fact, drugs that have been withdrawn from the market were shown to target highly variable genes [6]. To test this on our data, we got the list of the FDA-approved cancer drugs from the Genomics of Drug Sensitivity in Cancer (GDSC) database [25] and we checked the biological variability of their respective drug-target genes. We noticed that 99% of these drugs target genes are non-variable (Fig. 11A). Once we collected this list, we further annotated the genes as oncogene or tumor suppressor gene based on their oncogenic classification using OncoKb, a comprehensive and oncology-based database [26]. Out of these drug-target genes, 45% were oncogenes, 12% were tumor suppressor genes, and 2% had both oncogene/tumor suppression gene biological property (Fig. 11B) most of which are involved in apoptosis signaling, RTK signaling, EGFR signaling, and cell cycle regulation pathways.

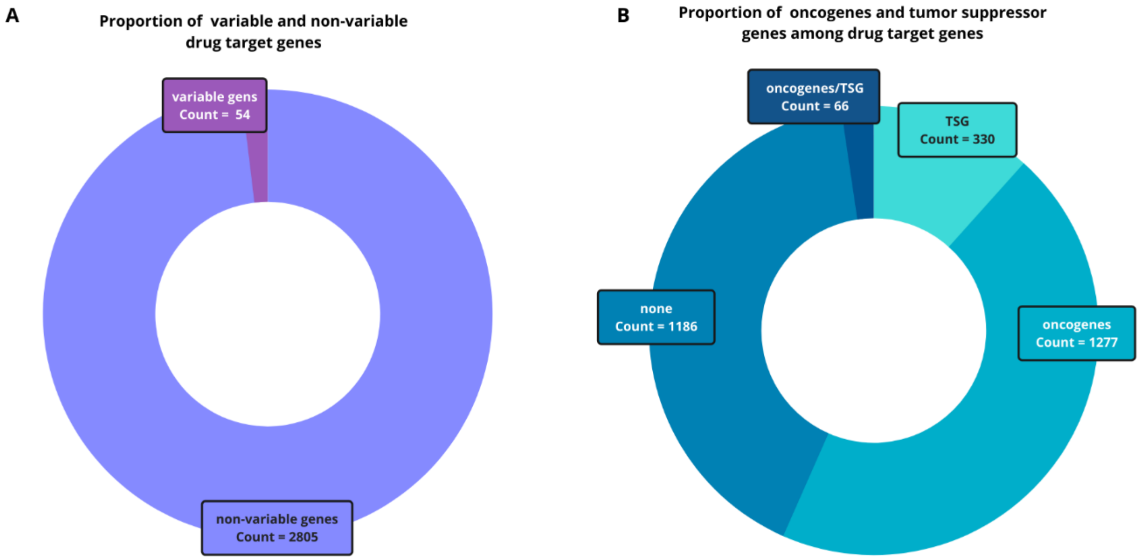

Figure 11. Donut plots showing the categorization of the drug-target genes from GDSC Database based on their variability score and oncogenic property. (A) Variable drug target genes (n=54) account for 1% of the total drug target genes that are targeted therapeutically by FDA approved drugs, while non-variable genes account (n=2805) account for 99%. **(B)** Stratification of the drug target genes based on their oncogenic property. Majority of the drug target genes are oncogenes (45%), while few have tumor suppressor genes (12%).

Since most of the approved drugs target non-variable genes with the above properties (Fig. 11), we collected the list of all the oncogenes and tumor suppressor genes from the 33 cancers and checked their respective biological variability score (Fig. 12). From the list, we identified MCL1 (MCL1 Apoptosis Regulator, BCL2 Family Member), a putative drug target oncogene, that possesses low variability score across several cancers (Fig. 12). This gene has shown to be frequently overexpressed in several human cancers and was directly linked to cancer drug resistance [51]. Several drugs targeting this gene have been released into the market like Obatoclax Mesylate, MIM1 [52][53]. This allowed us to look at other non-variable oncogenes like Eukaryotic initiation factor 4A2 (EIF4A2) and Calreticulin (CALR) which can serve as drug-target genes due to their low variability score and their involvement in cell-cycle regulation and apoptotic pathways giving them their oncogenic property [54][55].

Not only will this be important to derive the list of potential drug-target genes, but also to extract the ones that may elicit a varied drug response, like IGF2 and SERPINB3 (Fig. 12), considering that the responsiveness and effectiveness of drugs has been associated with expression variability [6]. Overall, this gives us a scope of all the potential genes that can be targeted therapeutically in each cancer based on their oncogenic property and biological variability score.

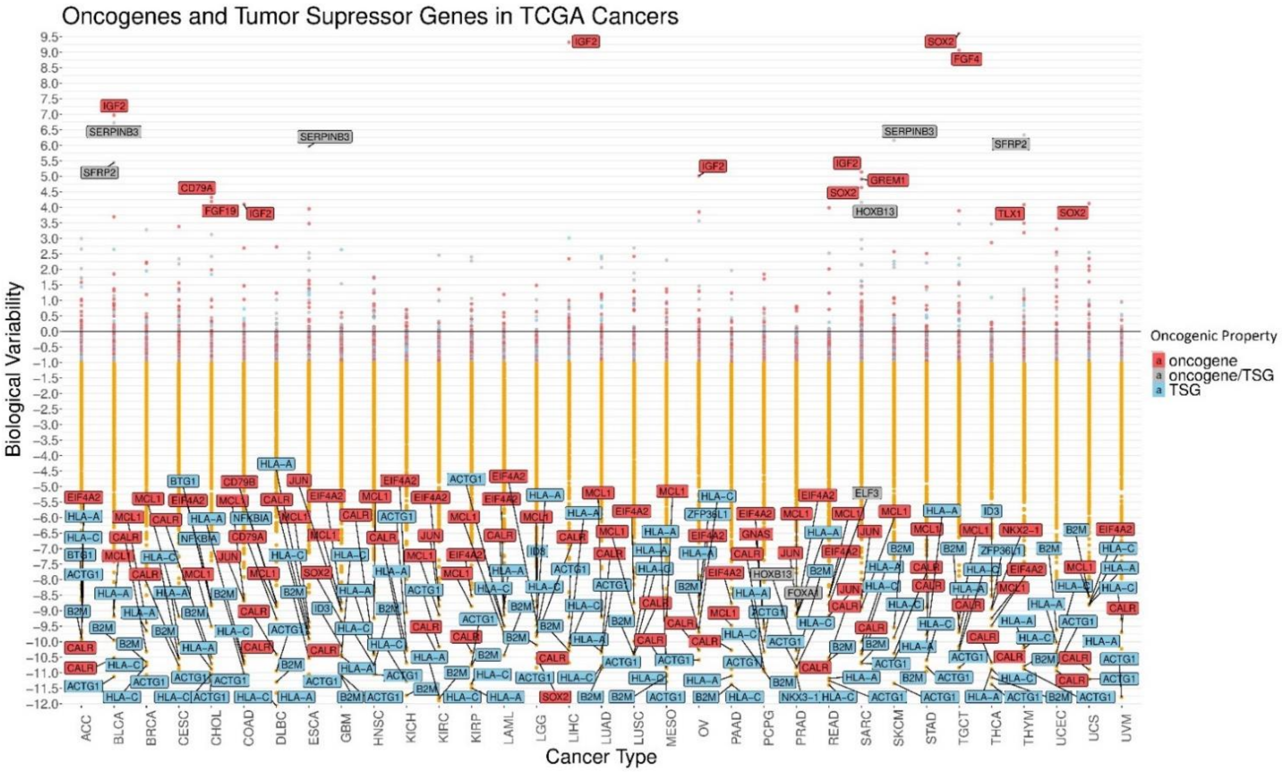

Figure 12.Potential oncogenes and tumor suppressor genes that can be targeted therapeutically. The blue boxes, gray and red boxes are oncogenes, onco/TSG, tumor

suppression genes respectively. The orange points represent all the possible drug-target genes.

CHAPTER V

DISCUSSION

Inter-individual variability is a fundamental issue in cancer research. It poses challenges in identifying molecular markers that can help us detect cancer at an early stage and predict response to drug treatment. This variability is the cumulative result of genetic alterations in expression patterns of key genes that play a role in disease progression and proliferation. In this study, we calculated biological variability using the RNA-seq data from the TCGA dataset for the purpose of identifying genes that play a significant role in cancer development and early detection.

Our first step was to choose the best normalization technique. As shown, log2(TPM +1) was the only normalization method unbiased towards different expression levels, unlike TPM, and it ranked better than DESeq vst transformation when answering biological questions. When vst transformed the data, it reduced the dependence of the variance on the mean, thus altering the gene expression data and consequently, the biological variability measure. However, log2(TPM +1) preserved the data and was the best performing normalization technique.

We applied our biological variability metric on 11,315 cases in 33 different cancers from the TCGA database. Once we obtained the list of non-variable and variable genes, we performed pan-cancer analysis, and we found out that physiologically related cancers have similar expression variability patterns. Moreover, we noticed that while some genes had similar biological variability scores between cancers, others were unique to their respective cancer. This suggests that we can extract mRNA-based molecular markers from biological variability scores. We linked diagnostic markers to non-variable genes since the former should have similar expression levels between patients and thus a low variability score. Conversely, we linked prognostic and tumor progression markers to highly variable genes since the tumor microenvironment changes as the patient progresses through cancer stages, leading to the downregulation or upregulation of genes in different pathological pathways. The results obtained were consistent with previous findings, as shown in the CIViC database, indicating that our metric is reliable.

Upon analyzing the biological variability of 198 cancer drug-target genes, we found out that all the FDA-approved drugs target non-variable genes. Response to drug treatment has been previously linked to expression variability, where drugs showed to be ineffective when targeting highly variable genes [6]. Once we annotated the drug-target genes, we noticed that most of them are non-variable oncogenes. This gave us the opportunity to search for all the oncogenes and tumor suppressor genes in our dataset, which can later be used as a reference for genes that can be targeted therapeutically.

Notably, there are some limitations to our study. Very few cancers in the TCGA data include clinical data on the drug treatment administered to the patients and treatment response and some cancers had missing information on tumor stage and vital state, which didn't allow us to dwell deeper into the relationship of gene expression variability and clinical response. However, we were able to use the GDDC database to get information on the drug-target genes and drug treatment. Future research should be done on detecting therapeutic biomarkers of drug sensitivity.

REFERENCES

1. Ashley, E.A. (2016). Towards precision medicine. Nature Reviews Genetics *17*, 507–522.

2. Meeks, J.J., Al-Ahmadie, H., Faltas, B.M., Taylor, J.A., Flaig, T.W., DeGraff, D.J., Christensen, E., Woolbright, B.L., McConkey, D.J., and Dyrskjøt, L. (2020). Genomic heterogeneity in bladder cancer: challenges and possible solutions to improve outcomes. Nature Reviews Urology *17*, 259–270.

3. Frerich, C.A., Brayer, K.J., Painter, B.M., Kang, H., Mitani, Y., El-Naggar, A.K., and Ness, S.A. (2017). Transcriptomes define distinct subgroups of salivary gland adenoid cystic carcinoma with different driver mutations and outcomes. Oncotarget *9*, 7341–7358.

4. H. Frederik Nijhout (2013). Stochastic Gene Expression: Dominance, Thresholds and Boundaries (Landes Bioscience).

5. Zuo, S., Zhang, X., and Wang, L. (2019). A RNA sequencing-based six-gene signature for survival prediction in patients with glioblastoma. Scientific Reports *9*.

6. Simonovsky, E., Schuster, R., and Yeger-Lotem, E. (2019). Large-scale analysis of human gene expression variability associates highly variable drug targets with lower drug effectiveness and safety. Bioinformatics *35*, 3028–3037.

7. Rajan, P., Stockley, J., Sudbery, I.M., Fleming, J.T., Hedley, A., Kalna, G., Sims, D., Ponting, C.P., Heger, A., Robson, C.N., et al. (2014). Identification of a candidate prognostic gene signature by transcriptome analysis of matched pre- and

post-treatment prostatic biopsies from patients with advanced prostate cancer. BMC Cancer *14*.

8. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics *45*, 1113–1120.
9. Zhang, J., and Huang, K. (2017). Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. BMC Genomics *18*.
10. Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell *158*, 929–944.
11. Cao, Z., and Zhang, S. (2016). An integrative and comparative study of pan-cancer transcriptomes reveals distinct cancer common and specific signatures. Scientific Reports *6*.
12. Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nature Reviews Genetics *20*, 631–656.
13. del Pino, M., Svanholm-Barrie, C., Torné, A., Marimon, L., Gaber, J., Sagasta, A., Persing, D.H., and Ordi, J. (2014). mRNA biomarker detection in liquid-based cytology: a new approach in the prevention of cervical cancer. Modern Pathology *28*, 312–320.
14. Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., et al. (2014).

Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes. Cell Reports *9*, 1235–1245.

15. Pennock, N.D., Jindal, S., Horton, W., Sun, D., Narasimhan, J., Carbone, L., Fei, S.S., Searles, R., Harrington, C.A., Burchard, J., et al. (2019). RNA-seq from archival FFPE breast cancer samples: molecular pathway fidelity and novel discovery. BMC Medical Genomics *12*.
16. Wang, S., Jia, M., He, Z., and Liu, X.-S. (2018). APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. Oncogene *37*, 3924–3936.
17. Wagner, G.P., Kin, K., and Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences *131*, 281–285.
18. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2009). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493–500.
19. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology *11*.
20. Bacher, R., Chu, L.-F., Leng, N., Gasch, A.P., Thomson, J.A., Stewart, R.M., Newton, M., and Kendziorski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. Nature Methods *14*, 584–586.

21. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2015). Ensembl 2016. Nucleic Acids Research *44*, D710–D716.
22. Kamel, H.F.M., and Al-Amodi, H.S.A.B. (2017). Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine. Genomics, Proteomics & Bioinformatics *15*, 220–235.
23. Lever, J., Jones, M.R., Danos, A.M., Krysiak, K., Bonakdar, M., Grewal, J.K., Culibrk, L., Griffith, O.L., Griffith, M., and Jones, S.J.M. (2019). Text-mining clinically relevant cancer biomarkers for curation into the CIViC database. Genome Medicine *11*.
24. Nalejska, E., Mączyńska, E., and Lewandowska, M.A. (2014). Prognostic and Predictive Biomarkers: Tools in Personalized Oncology. Molecular Diagnosis & Therapy *18*, 273–284.
25. Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research *41*, D955–D961.
26. Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology 1–16.
27. Quinn, T.P., Crowley, T.M., and Richardson, M.F. (2018). Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics *19*.

28. Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics *14*, 671–683.

29. Silander, O.K., Nikolic, N., Zaslaver, A., Bren, A., Kikoin, I., Alon, U., and Ackermann, M. (2012a). A Genome-Wide Analysis of Promoter-Mediated Phenotypic Noise in Escherichia coli. PLoS Genetics *8*, e1002443.

30. Alemu, E.Y., Carl, J.W., Corrada Bravo, H., and Hannenhalli, S. (2014). Determinants of expression variability. Nucleic Acids Research *42*, 3503–3514.

31. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research *18*, 1509–1517.

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

33. Hendrix, M.J.C., Seftor, E.A., Margaryan, N.V., and Seftor, R.E.B. (2017). Heterogeneity and Plasticity of Melanoma: Challenges of Current Therapies. Cutaneous Melanoma: Etiology and Therapy 57–65.

34. Jo, J., Choi, S., Oh, J., Lee, S.-G., Choi, S.Y., Kim, K.K., and Park, C. (2019). Conventionally used reference genes are not outstanding for normalization of gene expression in human cancer research. BMC Bioinformatics *20*.

35. Baran, B., Mert Ozupek, N., Yerli Tetik, N., Acar, E., Bekcioglu, O., and Baskin, Y. (2018). Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature. Gastroenterology Research *11*, 264–273.

36. Feng, H., Li, T., and Zhang, X. (2018). Characterization of kinase gene expression and splicing profile in prostate cancer with RNA-Seq data. BMC Genomics *19*.

37. Kim, H., and Kim, Y.-M. (2018a). Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types. Scientific Reports *8*.

38. Cao, Z., and Zhang, S. (2016). An integrative and comparative study of pan-cancer transcriptomes reveals distinct cancer common and specific signatures. Scientific Reports *6*.

39. Cabanski, C.R., White, N.M., Dang, H.X., Silva-Fisher, J.M., Rauck, C.E., Cicka, D., and Maher, C.A. (2015). Pan-cancer transcriptome analysis reveals long noncoding RNAs with conserved function. RNA Biology *12*, 628–642.

40. Penney, K.L., Schumacher, F.R., Kraft, P., Mucci, L.A., Sesso, H.D., Ma, J., Niu, Y., Cheong, J.K., Hunter, D.J., Stampfer, M.J., et al. (2011). Association of KLK3 (PSA) genetic variants with prostate cancer risk and PSA levels. Carcinogenesis *32*, 853–859.

41. Jelski, W., and Mroczko, B. (2020). Biochemical Markers of Colorectal Cancer – Present and Future<. Cancer Management and Research Volume 12, 4789–4797.

42. Savas, S., Liu, G., and Xu, W. (2013). Special considerations in prognostic research in cancer involving genetic polymorphisms. BMC Medicine 11.

43. Goossens, N., Nakagawa, S., Sun, X., and Hoshida, Y. (2015). Cancer biomarker discovery and validation. Translational Cancer Research *4*, 256–269.

44. Jayanthi, V.S.P.K.S.A., Das, A.B., and Saxena, U. (2020). Grade-specific diagnostic and prognostic biomarkers in breast cancer. Genomics *112*, 388–396.

45. Leivo, M.Z., Elson, P.J., Tacha, D.E., Delahunt, B., and Hansel, D.E. (2016). A combination of p40, GATA-3 and uroplakin II shows utility in the diagnosis and prognosis of muscle-invasive urothelial carcinoma. Pathology *48*, 543–549.

46. Kim, J.-S., Kim, E.S., Liu, D., Lee, J.J., Solis, L., Behrens, C., Lippman, S.M., Hong, W.K., Wistuba, I.I., and Lee, H.-Y. (2014). Prognostic Implications of Tumoral Expression of Insulin Like Growth Factors 1 and 2 in Patients With Non–Small-Cell Lung Cancer. Clinical Lung Cancer *15*, 213–221.

47. Ye, Y., Xiao, L., Wang, S.-J., Yue, W., Yin, Q.-S., Sun, M.-Y., Xia, W., Shao, Z.-Y., and Zhang, H. (2015b). Up-regulation of REG3A in colorectal cancer cells confers proliferation and correlates with colorectal cancer risk. Oncotarget *7*, 3921–3933.

48. Urbaniak, A., Jablonska, K., Podhorska-Okolow, M., Ugorski, M., and Dziegiel, P. (2018). Prolactin-induced protein (PIP)-characterization and role in breast cancer progression. American Journal of Cancer Research *8*, 2150–2164.

49. Martinez-Quetglas, I., Pinyol, R., Dauch, D., Torrecilla, S., Tovar, V., Moeini, A., Alsinet, C., Portela, A., Rodriguez-Carunchio, L., Solé, M., et al. (2016). IGF2 Is Up-regulated by Epigenetic Mechanisms in Hepatocellular Carcinomas and Is an Actionable Oncogene Product in Experimental Models. Gastroenterology *151*, 1192–1205.

50. Xiang, Q., Wu, W., Zhao, N., Li, C., Xu, J., Ma, L., Zhang, X., Xie, Q., Zhang, Z., Wang, J., et al. (2020). The influence of genetic polymorphisms in drug metabolism enzymes and transporters on the pharmacokinetics of different fluvastatin formulations. Asian Journal of Pharmaceutical Sciences *15*, 264–272.

51. Xiang, W., Yang, C.-Y., and Bai, L. (2018). MCL-1 inhibition in cancer treatment. OncoTargets and Therapy *Volume 11*, 7301–7314.

52. Sulkshane, P., and Teni, T. (2016). BH3 mimetic Obatoclax (GX15-070) mediates mitochondrial stress predominantly via MCL-1 inhibition and induces autophagy-dependent necroptosis in human oral cancer cells. Oncotarget *8*, 60060–60079.

53. Respondek, M., Beberok, A., Rok, J., Rzepka, Z., Wrześniok, D., and Buszman, E. (2018). MIM1, the Mcl-1 – specific BHgd3 mimetic induces apoptosis in human U87MG glioblastoma cells. Toxicology in Vitro *53*, 126–135.

54. Chen, Z.-H., Qi, J.-J., Wu, Q.-N., Lu, J.-H., Liu, Z.-X., Wang, Y., Hu, P.-S., Li, T., Lin, J.-F., Wu, X.-Y., et al. (2019). Eukaryotic initiation factor 4A2 promotes experimental metastasis and oxaliplatin resistance in colorectal cancer. Journal of Experimental & Clinical Cancer Research *38*.

55. Araki, M., and Komatsu, N. (2017). Novel molecular mechanism of cellular transformation by a mutant molecular chaperone in myeloproliferative neoplasms. Cancer Science *108*, 1907–19

APPENDIX

Cancer Type	Number of Diagnostic Markers	Number of Prognostic Markers
ACC	1	37
BLCA	0	58
BRCA	0	23
CESC	0	52
CHOL	0	62
COAD	3	13
DLBC	4	7
ESCA	0	107
GBM	13	0
HNSC	9	33
KICH	6	15
KIRC	1	12
KIRP	1	7
LAML	2	5
LGG	27	0
LIHC	30	44
LUAD	0	28
LUSC	0	26
MESO	0	17
OV	1	4
PAAD	0	34
PCPG	22	19
PRAD	11	4
READ	5	12
SARC	0	47
SKCM	1	78
STAD	0	38
TGCT	1	35
THCA	4	11
THYM	2	21
UCEC	0	28
UCS	0	39
UVM	6	4

Supplementary Table 1: Number of diagnostic and prognostic markers identified in each cancer.