International Journal of
Molecular Sciences

MDPI

Article

Pathway Activation Analysis for Pan-Cancer Personalized Characterization Based on Riemannian Manifold

Xingyi Li [1], Jun Hao [1], Junming Li [2], Zhelin Zhao [2], Xuequn Shang [1] and Min Li [3,*]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China; xingyili@nwpu.edu.cn (X.L.); haojun@mail.nwpu.edu.cn (J.H.); shang@nwpu.edu.cn (X.S.)
[2] School of Software, Northwestern Polytechnical University, Xi'an 710072, China; ljmsup@mail.nwpu.edu.cn (J.L.); zz263943@mail.nwpu.edu.cn (Z.Z.)
[3] School of Computer Science and Engineering, Central South University, Changsha 410083, China
* Correspondence: limin@mail.csu.edu.cn

Abstract: The pathogenesis of carcinoma is believed to come from the combined effect of polygenic variation, and the initiation and progression of malignant tumors are closely related to the dysregulation of biological pathways. Quantifying the alteration in pathway activation and identifying coordinated patterns of pathway dysfunction are the imperative part of understanding the malignancy process and distinguishing different tumor stages or clinical outcomes of individual patients. In this study, we have conducted in silico pathway activation analysis using Riemannian manifold (RiePath) toward pan-cancer personalized characterization, which is the first attempt to apply the Riemannian manifold theory to measure the extent of pathway dysregulation in individual patient on the tangent space of the Riemannian manifold. RiePath effectively integrates pathway and gene expression information, not only generating a relatively low-dimensional and biologically relevant representation, but also identifying a robust panel of biologically meaningful pathway signatures as biomarkers. The pan-cancer analysis across 16 cancer types reveals the capability of RiePath to evaluate pathway activation accurately and identify clinical outcome-related pathways. We believe that RiePath has the potential to provide new prospects in understanding the molecular mechanisms of complex diseases and may find broader applications in predicting biomarkers for other intricate diseases.

Keywords: pathway activation; Riemannian manifold; pan-cancer analysis; personalized characterization; pathway biomarkers

Citation: Li, X.; Hao, J.; Li, J.; Zhao, Z.; Shang, X.; Li, M. Pathway Activation Analysis for Pan-Cancer Personalized Characterization Based on Riemannian Manifold. *Int. J. Mol. Sci.* **2024**, *25*, 4411. https://doi.org/10.3390/ijms25084411

Academic Editor: Alexandre G. De Brevern

Received: 18 March 2024
Revised: 8 April 2024
Accepted: 11 April 2024
Published: 17 April 2024

1. Introduction

Carcinoma is driven by multiple factors and the underlying molecular mechanisms of cancer pathogenesis are complex; it is one of the most lethal diseases in the world. Genome-wide association studies (GWASs) and next-generation sequencing technologies [1] have continuously provided insights into the genetics of cancers, and varied single-gene biomarkers have been identified to play an important role in the early diagnosis, prognosis, and efficacy evaluation of cancers [2,3]. Canonically, most widely used methods are dedicated to finding differentially expressed genes [4,5]. However, the selection process is subjective, variations among samples are astronomical, and the functional understanding of the pathogenesis of carcinoma is intractable [6]. Meanwhile, with the heterogeneity of cells in tissues and the genetic heterogeneity between patients with complex diseases, most anticancer drugs are only effective in subgroups of patients [7]; it is urgent to develop personalized cancer treatments [8–10]. Pathway-based individualized analysis can overcome these challenges by using robust, aggregate features to reveal the molecular mechanisms of complex diseases [11].

Pathways are a series of biological activities among molecules in cells and are expressive of the biological processes within cells, such as metabolism, signaling, and growth

cycles, which will lead to some alteration or obtain some products [12]. Misregulation of activation for several important pathways has been found to be associated with cancer initiation and progression [13–16], identifying the pathways involved in the occurrence and progress of cancer and quantifying their dysregulation are the imperative part of understanding the process of malignancy. With the development of high-throughput technologies, a large amount of biological data has been generated, which has produced a rich set of pathway databases [17–21]. Therefore, it is of great significance to use bioinformatics tools to mine pathways related to the pathogenesis of cancers based on high-throughput data for the diagnosis and treatment.

Current quantitative pathway-centric measures have been proposed to evaluate the pathway activation and identify the dysregulated pathways in cancer initiation and progression. Some works extract critical features from expression values of genes belonging to pathways [10,22–24]. Some works consider the intrinsic structures of pathways; they calculate personalized pathway activation scores based on the topological information of pathways [11,25]. Many of these methods evaluate the pathway deregulation scores based on the Euclidean space, and although they are constantly improving the classification ability of diseases, they are still less than satisfactory in some analyses. They use Euclidean space since it is easy to be implemented and applied in practice. However, if the data samples do not lie on Euclidean space, the rationality and effectiveness of these methods cannot be guaranteed, since the differences calculated on the Euclidean space cannot represent the real geometrical relations among samples.

In order to further improve the classification accuracy, we have firstly developed a Riemannian manifold-based method, RiePath, to evaluate the pathway deregulation scores for each patient on the tangent space of the Riemannian manifold. It can not only convert gene-level expression information to pathway-level deregulation information, so as to achieve the dimensionality reduction, but also has the potential to identify essential biological pathways as biomarkers.

We compare RiePath to other feature engineering algorithms; the results show that our method can not only obtain higher clustering accuracy in the discrimination of normal and tumor samples and reproducibility, but also effectively capture the potential prognostic-related pathway biomarkers, which have the functional interpretability to explore the biological mechanism of carcinoma from a molecular level.

2. Results

2.1. Performance Comparison with Other Feature Engineering Methods

For further comparison, we first build a scheme from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/, accessed on 7 April 2024) Breast invasive carcinoma (BRCA) cohort to demonstrate the performance of our approach, as well as compare it with other state-of-the-art feature engineering methods. We first randomly select 50 samples from disease and normal samples. Then, the pathway activation matrix is calculated by each feature engineering algorithm according to the RNA-seq data of the randomly selected samples, where each row represents a pathway and each column represents a sample. Next, limma R package is used to perform differential expression analysis and select the top 10 differential expression pathways based on the adjusted p values. Subsequently, we utilize the hierarchical clustering algorithm to divide the 100 samples into two classes and compare the clustering results and sample labels through the adjusted rand index (ARI) evaluation index. The above steps are repeated 50 times in each method.

To demonstrate the discrimination effectiveness of pathway activation calculated by RiePath, We choose four representative pathway activation measurement algorithms: CORGs [10], GSVA [26], PLAGE [27], and ssGSEA [28], and the compared methods are implemented with default parameters.

CORGs defines a subset of genes in a pathway named as condition-responsive genes, which are considered to play a crucial role in each pathway. For each pathway, the pathway

activity score is evaluated by a combined z-score derived from the expression of condition-responsive genes.

GSVA is a gene set variation analysis method that calculates the variation in gene set enrichment over a sample population as the pathway activation. GSVA sorts genes based on the Kernel estimation of the cumulative density function of genes in the gene set, and calculates a Kolmogorov–Smirnov-like rank statistic for each gene set.

PLAGE can analyze the dysregulation level of a pathway by estimating the pathway activity based on the first eigenvector in the singular value decomposition of gene expression data.

ssGSEA is a single sample gene set enrichment analysis method, and the enrichment score is computed by the integration of the difference between weighted Empirical Cumulative Distribution Functions (ECDFs) of the genes in the signature and the ECDFs of the remaining genes.

The hierarchical clustering results measured by ARI index for clustering accuracy are shown in Figure 1. Higher ARI values means that samples with the same label are better clustered into one class, which can also prove that the feature engineering method of calculating the pathway activation can better capture the deviation of patients from the signature of healthy samples. As expected, the pathway dysregulation scores calculated by RiePath have more stable and better clustering performance than the other compared methods, and the pairwise comparison between RiePath and the other four methods using the *t* test can prove that the differences are statistically significant.

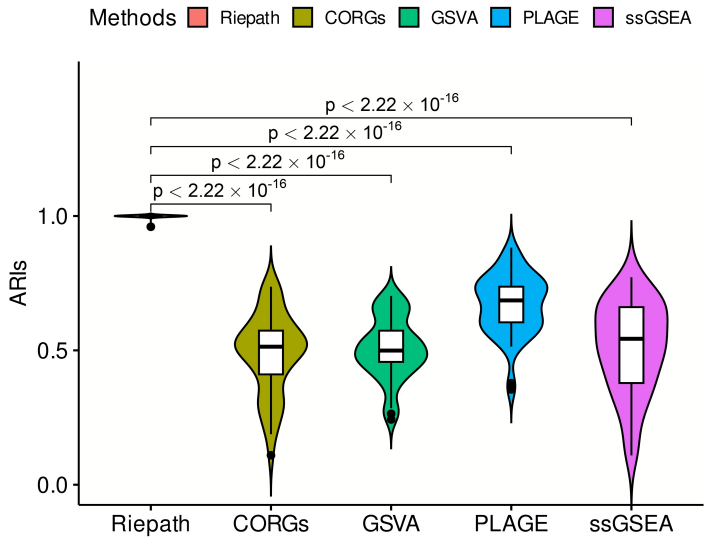

Figure 1. Hierarchical clustering comparison measured by ARI index.

Then, we compare the reproducibility of RNA-seq data to evaluate the how well gene-level sample differences are kept at the pathway level [12,29]. The reproducibility score (RS) is defined as the reciprocal of the weighted average of mean squared error (MSE):

$$RS = \frac{\binom{N}{2}}{\sum_{1 \le a,b \le N} \{Sim(X_a, X_b) - Sim(A_a, A_b)\}^2} \quad (1)$$

where $X = [X_1, \ldots, X_N]$ is the gene expression matrix, $A = [A_1, \ldots, A_N]$ is the pathway activation matrix, and Sim denotes the cosine similarity.

The smaller the distance between gene expression and pathway activation values, the greater the RS value is, which demonstrates that the sample space of the inferred pathway activation scores approaches that of the original gene expression. This is due to the fact that pathways only contain a small subset of genes, accounting for approximately one-third of all genes in the gene expression. After calculating the pathway activation scores, repeatability characteristics can measure how well the distance between samples

in terms of pathway activation can maintain the characteristics of the original data (i.e. all genes). As shown in Table 1, RiePath obtains the highest *RS* value compared to other feature engineering methods on each cancer type dataset, which means that our method is effective in retaining the characteristics of the original transcriptome data.

Table 1. Reproducibility comparison measured based on MSE for the five feature engineering algorithms.

Dataset	RiePath	CORGs	GSVA	PLAGE	ssGSEA
BLCA	198.78	0.98	1.04	0.98	196.87
BRCA	352.89	0.97	1.05	0.95	252.27
COAD	286.14	0.93	1.01	0.94	281.74
HNSC	263.76	0.95	1.05	0.97	254.19
KICH	389.52	0.83	1.02	0.77	293.59
KIRC	309.68	0.94	1.00	0.96	297.61
KIRP	387.01	0.91	1.01	0.94	316.69
LGG	688.92	0.94	1.00	0.91	635.00
LIHC	227.58	0.86	1.02	0.89	212.01
LUAD	271.53	0.97	1.03	0.95	242.96
LUSC	308.13	0.97	1.02	0.92	262.08
OV	408.63	0.92	1.02	0.94	358.85
PRAD	771.00	0.87	0.97	0.87	620.76
STAD	235.47	0.97	1.03	0.97	231.12
THCA	683.27	0.91	0.96	0.88	621.84
UCEC	220.73	0.98	1.07	0.97	217.65

2.2. The Identification of Dysregulated Pathways

For each pathway, all the tumor samples are firstly divided into two categories, dysregulated or near-normal samples, based on the mean and standard deviation of the normal samples' RiePath scores. If one tumor sample is classified as 'dysregulated' on this pathway, it means that the gene expression level of this tumor sample on this pathway severely deviates from that of the normal level, and this pathway is classified as 'dysregulated' on this tumor sample correspondingly. Otherwise, the tumor sample is classified as 'near-normal' on this pathway and this pathway is classified as 'near-normal' on this tumor sample accordingly.

For complex diseases like cancers, several vital pathways are often dysregulated simultaneously; the key to consider is how to effectively identify the dysregulated pathways for a specific tumor sample and whether there are differences among different cancer categories. Therefore, we comprehensively investigate the Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/, accessed on 7 April 2024) [17] pathways and Molecular Signatures Database (MSigDB, http://www.gsea-msigdb.org/gsea/msigdb, accessed on 7 April 2024) [30] Hallmark gene sets. Figure 2A shows the violin plot of percentage of dysregulated pathways on each sample in KEGG and Hallmark category; for the 16 tumor types, it can be seen that there are obvious differences. Kidney chromophobe (KICH) shows the highest proportion of dysregulated pathways, which is 98%. Conversely, Low-grade gliomas (LGG) shows the lowest proportion (79%). Interestingly, the percentages of dysregulated samples on each pathway in the KEGG and Hallmark category have less variation among the 16 kinds of cancer, as shown in Figure 2B. Moreover, for each of the 16 tumor categories, we also make a survey of the number of pathways dysregulated in all patients. As shown in Figure 2C, there are 344 pathways in the KEGG and Hallmark category, and the number of pathways perturbed in all disease samples are highest in KICH (256 pathways), and conversely, the least in LGG (32 pathways).

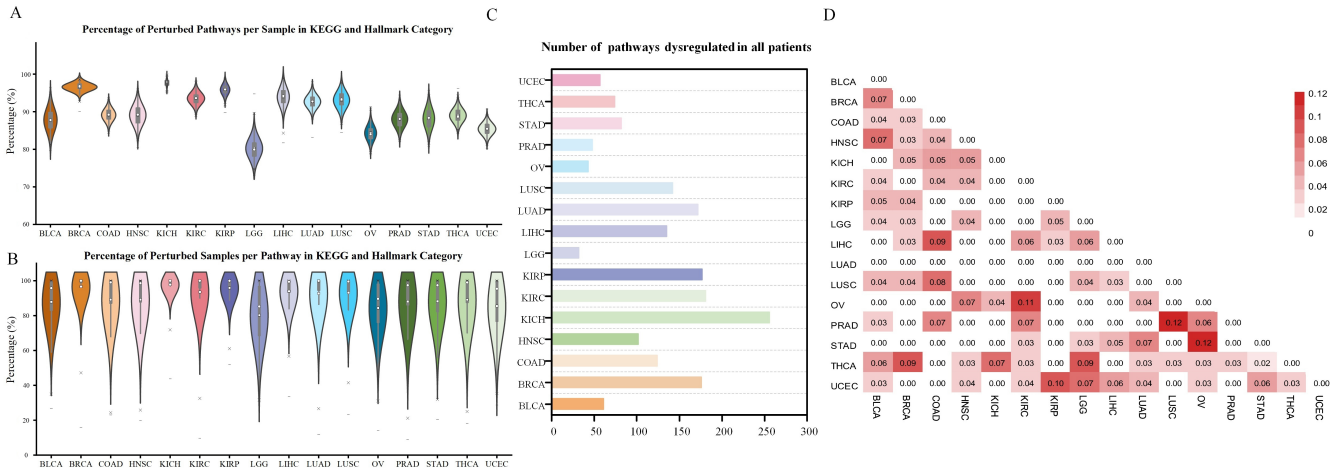

Figure 2. Survey of the important pathways in the 16 cancer types. (**A**) The violin plot of percentage of dysregulated pathways on each sample in KEGG and Hallmark category. (**B**) The violin plot of percentage of dysregulated samples on each pathway in KEGG and Hallmark category. (**C**) The number of pathways dysregulated in all patients. (**D**) The similarity of the overlapping prognostic pathway biomarkers identified in 16 cancer types using the Jaccard index.

2.3. The Identification of Prognostic Pathway Biomarkers

As the indicators to determine the progression and recurrence of tumors, prognostic biomarkers play an important role in cancer research. In this study, we propose RiePath to calculate the pathway activation of each pathway for each patient, and RiePath is applied in 344 pathways to 7291 disease samples across 16 TCGA cancer types. For each kind of cancer, we identify candidate prognostic pathway biomarkers using Kaplan–Meier analysis and setting the significance threshold of log-rank p value less than 0.05, which is consistent with the threshold used in many studies to identify prognostic biomarkers [7,31].

Supplementary Table S1 lists the prognostic pathway biomarkers identified based on the RiePath scores across the 16 cancer types. A total of 164 pathways are identified as promising prognostic pathway biomarkers across 16 cancer types, and the number of prognostic pathway biomarkers among different kinds of cancer are imbalanced. Most of the important pathways identified based on our method are mainly concentrated in Liver infiltrate hepatocellular carcinoma (LIHC) and Thyroid carcinoma (THCA), while a small number occur in the two kidney cancer types: KICH and Kidney renal papillary cell carcinoma (KIRP). Moreover, we introduce the Jaccard index to measure the similarity of the overlapping biomarkers identified in 16 cancer types from a pan-cancer perspective, and the Jaccard similarity index of two sets of prognostic pathways for each pair of cancer types is defined as follows:

$$J(u,v) = \frac{|u \cap v|}{|u \cup v|} = \frac{|u \cap v|}{|u| + |v| - |u \cap v|} \tag{2}$$

where u and v are the prognostic pathway biomarker sets of every two cancer types. If u and v are the same set, that is, the set of prognostic pathway biomarkers for the same cancer type, $J(u,v) = 0$. We observe that the overlapping prognostic pathway biomarkers shared by two cancer types in the KEGG and Hallmark category are very few (Figure 2D), meaning that the majority of prognostic pathway biomarkers identified in each cancer type are specific rather than shared, reflecting the diversity of human malignancies. This result is consistent with the conclusion presented in [7].

2.4. The Selection of Prognostic Pathway Biomarkers

Among the prognostic pathway biomarkers identified based on RiePath, there are many pathways closely related to the occurrence and progression of cancer, especially

signaling pathways. Mutations in cancer cell genomes affect signaling pathways that play key roles in cell growth, proliferation, angiogenesis, survival, apoptosis, and metastasis. Activation of these pathways will result in the upregulation of transcription factors that induce epithelial–mesenchymal transition in cells [32]. Several signaling pathways are critical for the embryonic development, which plays a key role in tumor progression and changes in response to the therapy in different cancers [33].

In BRCA, several signaling pathways are identified as prognostic pathway biomarkers based on RiePath scores from the KEGG database. The identification of the cAMP signaling pathway, PI3K/AKT signaling pathway, and VEGF signaling pathway have been verified by many studies to play an important role in the occurrence and development of breast cancer and are closely related to endocrine therapy resistance in the later period of breast cancer [34–38]. For example, it has been recognized that the elevated levels of intracellular cAMP will stimulate the growth of the normal human breast epithelial cells in culture [39]. Several studies have demonstrated that cAMP inhibits the growth of established breast cancer cell lines and breast cancer cells in primary culture [39–41]. In addition to the KEGG pathway database, we also identify HALLMARK_MTORC1_SIGNALING as the biomarker for BRCA. MTORC1 signaling has been supported in a previous study where *PIK3CA* mutations are associated with gene signature of low MTORC1 signaling in estrogen receptor-positive breast cancer, and *PIK3CA* mutations are one of the most common genetic aberrations in breast cancer [37,42].

Subsequently, we analyze two signaling pathways: "cAMP signaling pathway" and "HALLMARK_MTORC1_SIGNALING". From the waterfall plots and density plots in Figure 3A,B (first column shows the waterfall plots and second column shows the density plots), there is a significant difference between the overall RiePath values of the disease samples and that of the normal samples. Meanwhile, the *log*-rank test in the survival analysis (third column of Figure 3A,B) shows significance (p value < 0.05) on these two pathways, which indicates that these are two prognostic pathway biomarkers. Other visual summaries of signaling pathways identified by RiePath scores in the remaining cancer types are shown in Supplementary Figures S1–S15.

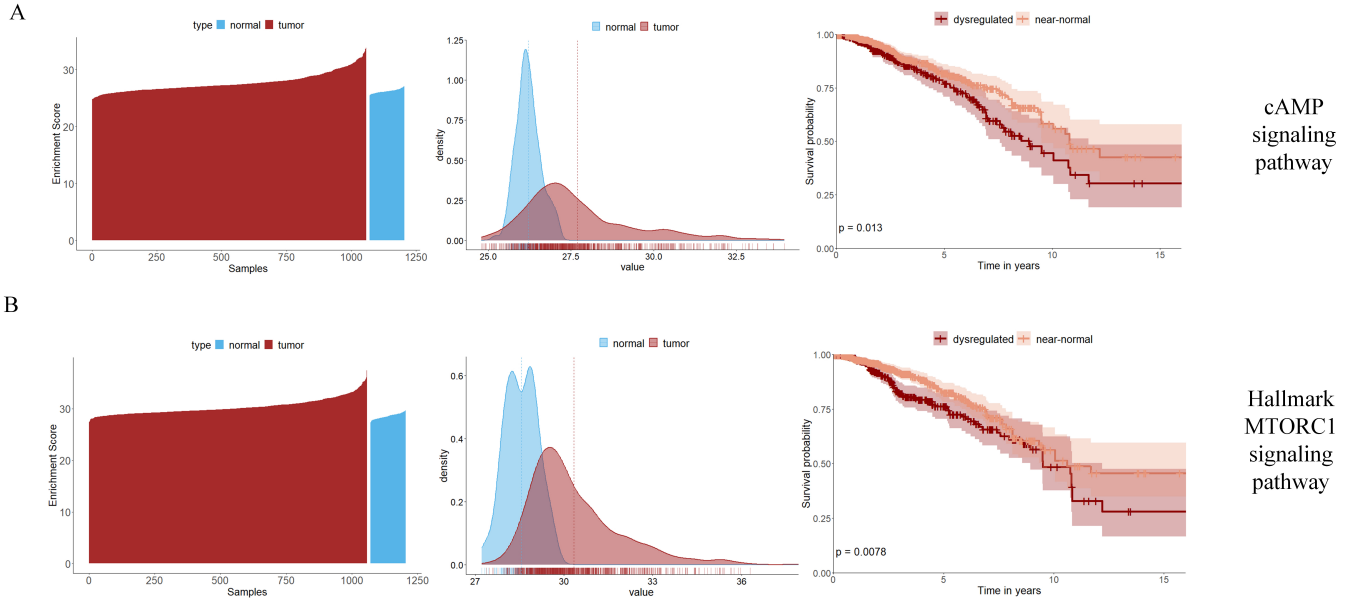

Figure 3. cAMP and Hallmark MTORC1 signaling pathways identified by RiePath scores in BRCA. (**A**) cAMP signaling pathway. (**B**) Hallmark MTORC1 signaling pathway. The waterfall and density plots of RiePath scores in tumor and normal samples are shown in the first and second columns, and the Kaplan–Meier plots indicate the significant survival difference for the dysregulated and near-normal patients in the two pathways.

3. Discussion

Gene-level information has been widely used in the studies of complex diseases, especially in cancer research. But they are often sensitive to noise and low repeatability. Integrating gene expression and pathway information can obtain an aggregate and biologically relevant representation, which quantifies the dysregulation of pathways and has the potential to identify essential pathways involved in the complex diseases.

In this study, we have proposed a Riemannian manifold-based method, RiePath, to evaluate the pathway dysregulation scores for each individual patient on the tangent space of the Riemannian manifold. RiePath converts high-dimensional gene-level expression information into relatively low-dimensional pathway-level dysregulation information, generating a compact and biologically relevant representation. This is the first attempt to apply Riemannian manifold theory to infer the degree of dysregulation at the pathway level so as to understand the molecular mechanisms of diseases. We apply the proposed method to the analysis of 16 cancer types. The results demonstrate that RiePath can not only have higher performance, but also effectively capture the potential prognostic pathway biomarkers with functional interpretability, which can not only provide better understandings of the mechanisms of cancer progression and drug therapy, but is also critical to the improvement of treatments.

RiePath is a clinically relevant predictive method that measures pathway dysregulation scores using pathway and gene expression data. It can also be extended to any other kind of data with known pathway assignments. Meanwhile, the development of single-cell sequencing technologies provides the possibility to reveal the mechanism of complex diseases at the cellular level [43,44]. RiePath is data-based and the pathway dysregulation is context-specific, so it also has the potential to be extended to single-cell study in future work.

4. Materials and Methods

4.1. Data Collection

Gene expression profiles and corresponding clinical information used in this study are generated by TCGA datasets. The high-throughput sequencing data of TCGA are downloaded from the UCSC Xena browser (https://xenabrowser.net/, accessed on 7 April 2024), and we obtain the Fragments per Kilobase of transcript per Million mapped reads (FPKM) processed normalized gene expression profiles from each of cancer types. All the data are $log2$ transformed. We only keep the first vital in a sequence of samples. Considering that RiePath and some comparison algorithms need to use normal samples as the background group, we download the tumor-adjacent tissue samples as normal samples. Meanwhile, cancer types with more than 10 normal samples and disease samples will be used in the pan-cancer study to ensure the sufficient sample sizes. Totally, 16 TCGA projects (BLCA, BRCA, COAD, LGG, HNSC, PRAD, THCA, KIRC, KIRP, KICH, LIHC, LUAD, LUSC, OV, STAD, UCEC) meet the requirements, as shown in Table 2 and Figure 4B,C.

Pathways come from the KEGG database and MSigDB Hallmark gene sets. For the pathway data from KEGG database, the KGML (KEGG XML) files are parsed into graph models maintaining all essential pathway attributes using the KEGGgraph package [45]. For a pathway, it contains nodes and edges; a node represents a gene, and an edge is the interaction between two genes. Considering the subsequent analysis, genes not existing in the RNA-seq gene expression data, and the corresponding edges, are discarded. Therefore, 294 pathways with 4156 nodes and 17,349 edges are obtained. Meanwhile, 50 biological pathways from the MSigDB Hallmark signature collection are also considered into this study. In total, there are 344 pathways used in our analysis.

Table 2. Summary of TCGA datasets for pathway analysis.

Code	Source	Number of Tumor Samples	Number of Normal Samples	Number of Total Samples
BLCA	Bladder carcinoma	400	23	423
BRCA	Breast invasive carcinoma	1057	136	1193
COAD	Colorectal adenocarcinoma	432	54	486
LGG	Low-grade gliomas	509	14	523
HNSC	Head and neck squamous cell carcinoma	496	48	544
PRAD	Prostate adenocarcinoma	428	68	496
THCA	Thyroid carcinoma	504	62	566
KIRC	Kidney renal clear cell carcinoma	523	78	601
KIRP	Kidney renal papillary cell carcinoma	285	32	317
KICH	Kidney chromophobe	64	23	87
LIHC	Liver infiltrate hepatocellular carcinoma	36	5	417
LUAD	Lung adenocarcinoma	499	72	571
LUSC	Lung squamous cell carcinoma	489	52	541
OV	Ovarian carcinoma	358	19	377
STAD	Stomach adenocarcinoma	348	32	380
UCEC	Endometrioid carcinoma	534	32	566

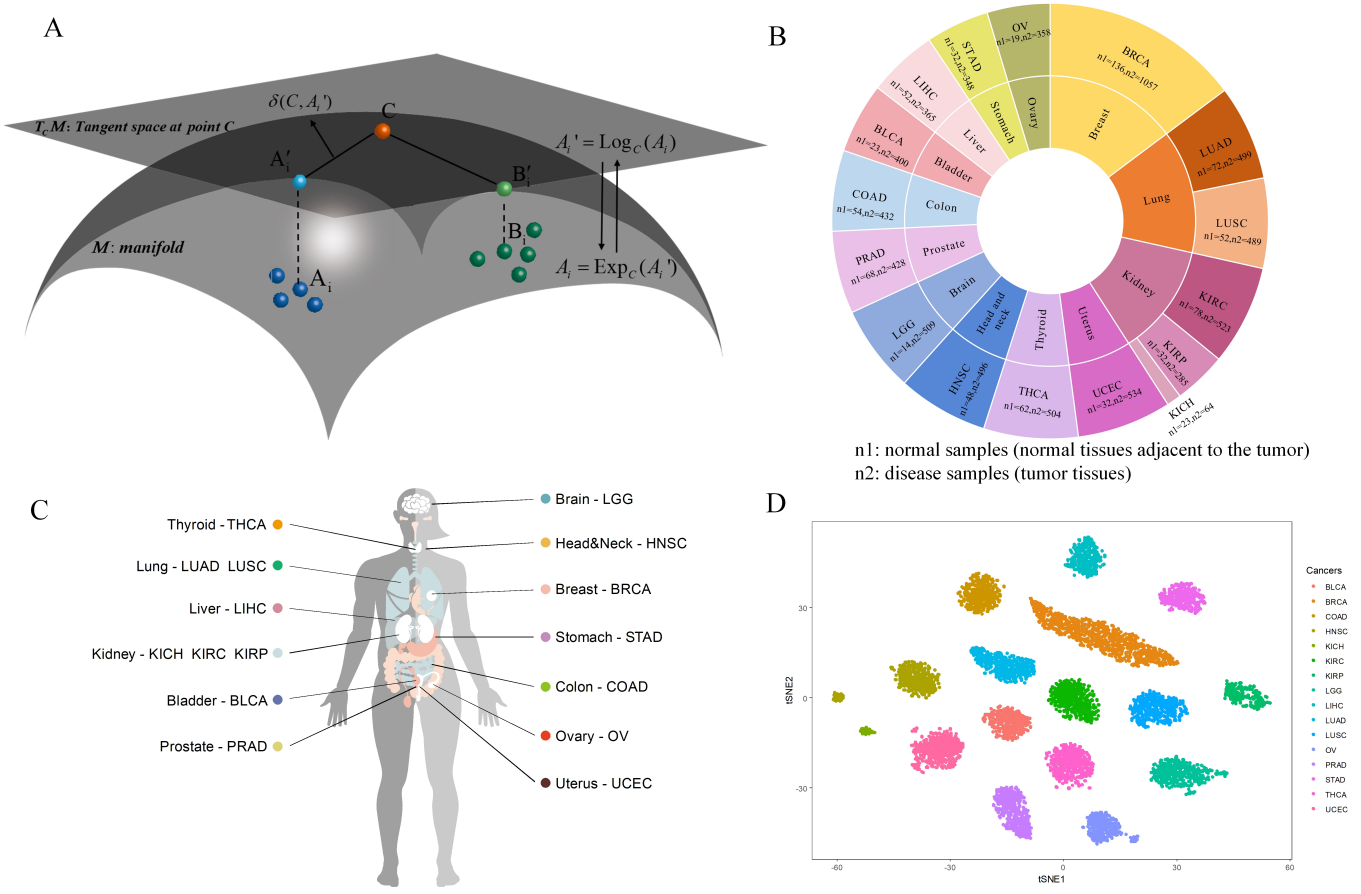

Figure 4. Overview of the RiePath algorithm. (**A**) Illustration of the Riemannian manifold and tangent space at point C. (**B**) The number of samples in each cancer type. (**C**) The 16 TCGA cancer types that are analyzed in the pan-cancer study. (**D**) t-SNE data visualization of the RiePath scores from all patients with tumor tissues of the 16 cancer types.

4.2. Space of Symmetric Positive Definite (SPD) Matrices

Given a gene expression matrix $X_i \in \mathbb{R}^{N \times E}$ with N genes and E samples, the space of symmetric matrices can be denoted as follows:

$$S(N) = \{S \in R^{N \times N}, S^T = S\} \tag{3}$$

and the space of positive-definite matrices can be denoted as:

$$P(N) = \{P \in R^{N \times N}, u^T P u > 0, \forall u \in R^N\} \tag{4}$$

the space of SPD matrices is denoted as:

$$SPD(N) = S(N) \cap P(N) \tag{5}$$

The space of SPD matrices lie on a differentiable Riemannian manifold M (Figure 4A) with the dimensionality of $N(N+1)/2$ [46].

For two spatial covariance matrices (two points) $A_i \in \mathbb{R}^{N \times N}$ and $C \in \mathbb{R}^{N \times N}$ on the manifold, the Riemannian distance $\delta_R(C, A_i)$ between them is defined as:

$$\delta_R(C, A_i) = \left\| log(C^{-1} A_i) \right\|_F = \sqrt{\sum_{n=1}^{N} log^2(\lambda_n)} \tag{6}$$

where $log(\cdot)$ denotes the logarithmic operator, $|| \cdot ||_F$ is the Frobenius norm of a matrix, and $\{\lambda_n\}_{n=1}^N$ the real strictly positive eigenvalues of $C^{-1} A_i$. This distance represents the length of unique shortest curve (called geodesic) connecting the two points in the Riemannian manifold [47].

4.3. Riemannian Tangent Space

The derivatives of each point in the Riemannian manifold can form a tangent space $T_C M$, and it consists of a set of tangent vectors [48,49]. The tangent space, as a Euclidean space, is an important space in the analysis of a Riemannian manifold. The logarithmic mapping operator that can project a point on the manifold to the tangent space at point C is defined as:

$$A_i{}' = Log_C(A_i) = C^{\frac{1}{2}} log(C^{-\frac{1}{2}} A_i C^{-\frac{1}{2}}) C^{\frac{1}{2}} \tag{7}$$

The inverse operation is the exponential mapping that projects a tangent element back to the original manifold:

$$A_i = Exp_C(A_i{}') = C^{\frac{1}{2}} exp(C^{-\frac{1}{2}} A_i{}' C^{-\frac{1}{2}}) C^{\frac{1}{2}} \tag{8}$$

where $log(\cdot)$ and $exp(\cdot)$ are the logarithm and exponential of a matrix, respectively. They are operators that map one-to-one between Riemannian manifold and tangent space. Figure 4A illustrates this process.

At each point of the manifold, a scalar product can be defined in the associated tangent space. The tangent space, as a Euclidean space, is an important space in the analysis of a Riemannian manifold. The tangent space $T_c M = \{Log_C(A_i), A_i \in SPD(N)\}$ at point C is a space of symmetric matrices and there are only $N(N+1)/2$ independent elements. The minimal representation of the tangent space can be found as a vector space [50]:

$$T_c M = \{A_i{}' = upper(C^{-\frac{1}{2}} Log_C(A_i) C^{-\frac{1}{2}}) \in R^{N(N+1)/2}\} \tag{9}$$

where the operator $upper(\cdot)$ is to keep the upper triangular portion of symmetric matrix and vectorize it.

The distance between point C and point A_i on the Riemannian manifold M can be measured as the Euclidean distance from point C and point A_i' the tangent space at point C [49]:

$$\delta(C, A_i') = ||A_i' - 0||_2 \quad (10)$$

where $A_i' \in T_C M$ the vector in tangent space corresponding to $A_i \in M$. The tangent space $T_C M$ of Riemannian manifold M at point C is shown in Figure 4A.

4.4. Pathway Dysregulation Scores

RiePath is a novel computational algorithm to evaluate pathway deregulation scores for each individual sample based on Riemannian manifold, enabling the accurate evaluation of pathway activation and the identification of the powerful pathway biomarkers for the pan-cancer personalized characterization across 16 cancer types.

RiePath aligns pathways and gene expression data to project samples into the space of SPD matrices, which evaluates the degree of pathway dysregulation from normal to disease. To estimate the pathway dysregulation score, we utilize the expression values of genes belonging to each pathway. There are three steps to measure the pathway deregulation scores: (1) The covariance matrix of the group of normal samples is calculated. The covariance matrix is an SPD matrix, which lies on a Riemannian manifold and will be projected as a point, that is, the reference point C in Figure 4A. (2) An individual patient will be added into the group of healthy samples, and we will obtain a point whose distance from the reference point represents the perturbation of the patient. Repeat this step until we obtain all the perturbed points in the Riemannian manifold based on all the patients. (3) We map the perturbed points in the Riemannian manifold to the tangent space at the reference point C and calculate the distance between reference point and each perturbed point. This distance is considered to be the extent to which each pathway deviates from the normal in each patient, that is, the extent of dysregulation. After calculating the distances between all perturbed points and reference point for all the pathways, the high-dimensional gene expression matrix will be converted into a low-dimensional matrix for pathways and patients. The t-SNE data visualization of RiePath scores for all tumor samples across the 16 cancer types is shown in Figure 4D. It indicates that tumor samples from the same cancer type are compactly clustered together and different types of tumor samples are separated from each other well, which means that the pathway activation values obtained by our algorithm can effectively distinguish samples of different cancer types. The pathway activation analysis can provide effective assistance in identifying cancer-specific pathway biomarkers for precision medicine.

5. Conclusions

In this study, we test a novel method named RiePath, which analyzes the pathway activation and identifies coordinated patterns of pathway dysregulation using Riemannian manifold on pan-cancer data. Unlike most of the existing pathway-based inference tools that calculate the activation of pathways in the Euclidean space, we attempt to introduce a Riemannian manifold-based method to evaluate the pathway activation for each disease sample on the tangent space of the Riemannian manifold. We compare the performance of RiePath and some other feature engineering algorithms, and identify the dysregulated pathways and candidate prognostic pathway biomarkers based on the RiePath scores. The results prove the effectiveness of introducing Riemannian manifold to evaluate the personalized pathway activation for pan-cancer analysis, the effectiveness of generating a relatively low-dimensional and biologically relevant representation, and the robustness of identifying a panel of biologically meaningful pathway signatures as biomarkers. We believe that RiePath has the potential to provide new prospects in understanding the molecular mechanisms of complex diseases and may find broader applications in predicting pathway biomarkers for other intricate diseases.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms25084411/s1.

Author Contributions: Conceived and designed the experiments: X.L. and M.L. Performed the experiments and analyzed the data: X.L., J.H., J.L. and Z.Z. Wrote the paper: X.L. and J.H. Designed the software used in analysis: X.L. and X.S. Oversaw the project: M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant No. 62202383, the National Key Research and Development Program of China under Grant No. 2022YFD1801200.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article and Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, X.; Li, Y.; Shang, X.; Kong, H. A sequence-based machine learning model for predicting antigenic distance for H3N2 influenza virus. *Front. Microbiol.* **2024**, *15*, 1345794. [CrossRef] [PubMed]
2. Li, X.; Xiang, J.; Wu, F.X.; Li, M. A dual ranking algorithm based on the multiplex network for heterogeneous complex disease analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *19*, 1993–2002. [CrossRef] [PubMed]
3. Li, X.; Xiang, J.; Wang, J.; Li, J.; Wu, F.X.; Li, M. FUNMarker: Fusion network-based method to identify prognostic and heterogeneous breast cancer biomarkers. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *18*, 2483–2491. [CrossRef] [PubMed]
4. Rapaport, F.; Khanin, R.; Liang, Y.; Pirun, M.; Krek, A.; Zumbo, P.; Mason, C.E.; Socci, N.D.; Betel, D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **2013**, *14*, 1–13. [CrossRef] [PubMed]
5. Soneson, C.; Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* **2013**, *14*, 91. [CrossRef] [PubMed]
6. Goel, G.; Conway, K.L.; Jaeger, M.; Netea, M.G.; Xavier, R.J. Multivariate inference of pathway activity in host immunity and response to therapeutics. *Nucleic Acids Res.* **2014**, *42*, 10288–10306. [CrossRef] [PubMed]
7. Uhlen, M.; Zhang, C.; Lee, S.; Sjöstedt, E.; Fagerberg, L.; Bidkhori, G.; Benfeitas, R.; Arif, M.; Liu, Z.; Edfors, F.; et al. A pathology atlas of the human cancer transcriptome. *Science* **2017**, *357*, eaan2507. [CrossRef] [PubMed]
8. Symmans, W.F.; Liu, J.; Knowles, D.M.; Inghirami, G. Breast cancer heterogeneity: Evaluation of clonality in primary and metastatic lesions. *Hum. Pathol.* **1995**, *26*, 210–216. [CrossRef] [PubMed]
9. Ein-Dor, L.; Kela, I.; Getz, G.; Givol, D.; Domany, E. Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics* **2005**, *21*, 171–178. [CrossRef]
10. Lee, E.; Chuang, H.Y.; Kim, J.W.; Ideker, T.; Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* **2008**, *4*, e1000217. [CrossRef]
11. Li, X.; Li, M.; Zheng, R.; Chen, X.; Xiang, J.; Wu, F.X.; Wang, J. Evaluation of pathway activation for a single sample toward inflammatory bowel disease classification. *Front. Genet.* **2020**, *10*, 1401. [CrossRef]
12. Lim, S.; Lee, S.; Jung, I.; Rhee, S.; Kim, S. Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Brief. Bioinform.* **2020**, *21*, 36–46. [CrossRef]
13. Drier, Y.; Sheffer, M.; Domany, E. Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 6388–6393. [CrossRef]
14. Mao, W.; Zaslavsky, E.; Hartmann, B.M.; Sealfon, S.C.; Chikina, M. Pathway-level information extractor (PLIER) for gene expression data. *Nat. Methods* **2019**, *16*, 607–610. [CrossRef]
15. Li, F.; Wu, T.; Xu, Y.; Dong, Q.; Xiao, J.; Xu, Y.; Li, Q.; Zhang, C.; Gao, J.; Liu, L.; et al. A comprehensive overview of oncogenic pathways in human cancer. *Brief. Bioinform.* **2020**, *21*, 957–969. [CrossRef]
16. Li, X.; Li, M.; Xiang, J.; Zhao, Z.; Shang, X. SEPA: Signaling entropy-based algorithm to evaluate personalized pathway activation for survival analysis on pan-cancer data. *Bioinformatics* **2022**, *38*, 2536–2543. [CrossRef]
17. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef]
18. Romero, P.; Wagg, J.; Green, M.L.; Kaiser, D.; Krummenacker, M.; Karp, P.D. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **2005**, *6*, R2. [CrossRef]
19. Pico, A.R.; Kelder, T.; Van Iersel, M.P.; Hanspers, K.; Conklin, B.R.; Evelo, C. WikiPathways: Pathway editing for the people. *PLoS Biol.* **2008**, *6*, e184. [CrossRef]
20. Schaefer, C.F.; Anthony, K.; Krupa, S.; Buchoff, J.; Day, M.; Hannay, T.; Buetow, K.H. PID: The pathway interaction database. *Nucleic Acids Res.* **2009**, *37*, D674–D679. [CrossRef]
21. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2020**, *48*, D498–D503. [CrossRef]

22. Huang, E.; Ishida, S.; Pittman, J.; Dressman, H.; Bild, A.; Kloos, M.; D'Amico, M.; Pestell, R.G.; West, M.; Nevins, J.R. Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.* **2003**, *34*, 226–230. [CrossRef]
23. Bild, A.H.; Yao, G.; Chang, J.T.; Wang, Q.; Potti, A.; Chasse, D.; Joshi, M.B.; Harpole, D.; Lancaster, J.M.; Berchuck, A.; et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **2006**, *439*, 353–357. [CrossRef]
24. Young, M.R.; Craft, D.L. Pathway-informed classification system (PICS) for cancer analysis using gene expression data. *Cancer Inform.* **2016**, *15*, CIN-S40088. [CrossRef]
25. Han, L.; Maciejewski, M.; Brockel, C.; Gordon, W.; Snapper, S.B.; Korzenik, J.R.; Afzelius, L.; Altman, R.B. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics* **2018**, *34*, 985–993. [CrossRef]
26. Hänzelmann, S.; Castelo, R.; Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **2013**, *14*, 7. [CrossRef]
27. Tomfohr, J.; Lu, J.; Kepler, T.B. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinform.* **2005**, *6*, 225. [CrossRef]
28. Barbie, D.A.; Tamayo, P.; Boehm, J.S.; Kim, S.Y.; Moody, S.E.; Dunn, I.F.; Schinzel, A.C.; Sandy, P.; Meylan, E.; Scholl, C.; et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **2009**, *462*, 108–112. [CrossRef]
29. Vitali, F.; Li, Q.; Schissler, A.G.; Berghout, J.; Kenost, C.; Lussier, Y.A. Developing a 'personalome' for precision medicine: Emerging methods that compute interpretable effect sizes from single-subject transcriptomes. *Brief. Bioinform.* **2019**, *20*, 789–805. [CrossRef]
30. Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J.P.; Tamayo, P. The molecular signatures database hallmark gene set collection. *Cell Syst.* **2015**, *1*, 417–425. [CrossRef]
31. Su, K.; Yu, Q.; Shen, R.; Sun, S.Y.; Moreno, C.S.; Li, X.; Qin, Z.S. Pan-cancer analysis of pathway-based gene expression pattern at the individual level reveals biomarkers of clinical prognosis. *Cell Rep. Methods* **2021**, *1*, 100050. [CrossRef] [PubMed]
32. Takebe, N.; Warren, R.Q.; Ivy, S.P. Breast cancer growth and metastasis: Interplay between cancer stem cells, embryonic signaling pathways and epithelial-to-mesenchymal transition. *Breast Cancer Res.* **2011**, *13*, 211. [CrossRef]
33. Schmid, S.; Bieber, M.; Zhang, F.; Zhang, M.; He, B.; Jablons, D.; Teng, N.N. Wnt and hedgehog gene pathway expression in serous ovarian cancer. *Int. J. Gynecol. Cancer* **2011**, *21*, 975. [CrossRef]
34. Dong, H.; Claffey, K.P.; Brocke, S.; Epstein, P.M. Inhibition of breast cancer cell migration by activation of cAMP signaling. *Breast Cancer Res. Treat.* **2015**, *152*, 17–28. [CrossRef]
35. Tang, X.; Zhang, Q.; Shi, S.; Yen, Y.; Li, X.; Zhang, Y.; Zhou, K.; Le, A.D. Bisphosphonates suppress insulin-like growth factor 1-induced angiogenesis via the HIF-1α/VEGF signaling pathways in human breast cancer cells. *Int. J. Cancer* **2010**, *126*, 90–103. [CrossRef]
36. Spangle, J.M.; Dreijerink, K.M.; Groner, A.C.; Cheng, H.; Ohlson, C.E.; Reyes, J.; Lin, C.Y.; Bradner, J.; Zhao, J.J.; Roberts, T.M.; et al. PI3K/AKT signaling regulates H3K4 methylation in breast cancer. *Cell Rep.* **2016**, *15*, 2692–2704. [CrossRef]
37. Madsen, R.R.; Erickson, E.C.; Rueda, O.M.; Robin, X.; Caldas, C.; Toker, A.; Semple, R.K.; Vanhaesebroeck, B. Positive correlation between transcriptomic stemness and PI3K/AKT/mTOR signaling scores in breast cancer, and a counterintuitive relationship with PIK3CA genotype. *PLoS Genet.* **2021**, *17*, e1009876. [CrossRef]
38. Zhu, K.; Wu, Y.; He, P.; Fan, Y.; Zhong, X.; Zheng, H.; Luo, T. PI3K/AKT/mTOR-targeted therapy for breast cancer. *Cells* **2022**, *11*, 2508. [CrossRef] [PubMed]
39. Starzec, A.B.; Spanakis, E.; Nehme, A.; Salle, V.; Veber, N.; Mainguene, C.; Planchon, P.; Valette, A.; Prevost, G.; Israel, L. Proliferative responses of epithelial cells to 8-bromo-cyclic AMP and to a phorbol ester change during breast Pathogenesis. *J. Cell. Physiol.* **1994**, *161*, 31–38. [CrossRef]
40. Cho-Chung, Y.S. Suppression of malignancy targeting cyclic AMP signal transducing proteins. *Biochem. Soc. Trans.* **1992**, *20*, 425–430. [CrossRef]
41. Kim, S.N.; Ahn, Y.H.; Kim, S.G.; Park, S.D.; Cho-Chung, Y.S.; Hong, S.H. 8-Cl-cAMP induces cell cycle-specific apoptosis in human cancer cells. *Int. J. Cancer* **2001**, *93*, 33–41. [CrossRef] [PubMed]
42. Loi, S.; Haibe-Kains, B.; Majjaj, S.; Lallemand, F.; Durbecq, V.; Larsimont, D.; Gonzalez-Angulo, A.M.; Pusztai, L.; Symmans, W.F.; Bardelli, A.; et al. PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor–positive breast cancer. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 10208–10213. [CrossRef] [PubMed]
43. Teschendorff, A.E.; Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat. Commun.* **2017**, *8*, 15599. [CrossRef] [PubMed]
44. Hu, J.; Chen, M.; Zhou, X. Effective and scalable single-cell data alignment with non-linear canonical correlation analysis. *Nucleic Acids Res.* **2022**, *50*, e21. [CrossRef] [PubMed]
45. Zhang, J.D.; Wiemann, S. KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **2009**, *25*, 1470–1471. [CrossRef] [PubMed]
46. Förstner, W.; Moonen, B. A metric for covariance matrices. In *Geodesy—The Challenge of the 3rd Millennium*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 299–309.
47. Congedo, M.; Barachant, A.; Bhatia, R. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Comput. Interfaces* **2017**, *4*, 155–174. [CrossRef]
48. Nguyen, C.H.; Artemiadis, P. EEG feature descriptors and discriminant analysis under Riemannian Manifold perspective. *Neurocomputing* **2018**, *275*, 1871–1883. [CrossRef]

49. Barachant, A.; Bonnet, S.; Congedo, M.; Jutten, C. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Trans. Biomed. Eng.* **2011**, *59*, 920–928. [CrossRef]
50. Tuzel, O.; Porikli, F.; Meer, P. Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1713–1727. [CrossRef]