# Deep Variational Autoencoders and Muti-Omics Integration for Pancreatic Cancer Subtyping

*Aadil Rashid Bhat*[a]*, *Rana Hashmy*[a]

[a]*Department of Computer Sciences, University of Kashmir,190006, India*

**Abstract**

**in both men and women, which claimed close to half a million deaths in 2020. Characterized by late diagnosis, poor survival rates, and high incidence of metastasis, Pancreatic cancer is predicted to become the second leading cause of cancer-related deaths by 2030. However, there has been little advancement in terms of early detection and effective treatments for Pancreatic Cancer, leading to a dismal 5-year survival rate of 3-15%. Which renders the unmet challenge of early diagnosis of PC both urgent and important. Recently, multi-omics analysis of numerous cancers has provided a new perspective on genomics, epigenomics, and transcriptomics deregulations in cancer. Which helped with fine-tuned characterization, classification, and early diagnosis of Cancers. However, due to the vast number of heterogeneous variables in each omics data, and its disparate and dynamic nature multi-omics data possess many challenges in data integration and knowledge discovery. In this study, we used Deep learning to find a latent representation of integrated multi-omics data together with many clustering methods to find homogeneous pancreatic cancer subtypes. Which can explain differences in disease trajectories and outcomes in heterogeneous cohorts. And help improve early diagnosis, treatment, and prognosis of pancreatic cancer.**

*Keywords: Multi-Omics; Data Integration; Deep Variational Autoencoder; Cancer Subtyping.*

## 1. Introduction

Pancreatic cancer is a highly lethal form of cancer, ranking as the seventh most common cause of cancer-related fatalities worldwide. [1]. Pancreatic ductal adenocarcinoma (PDAC) has a very high mortality rate and its incidence is increasing, it has the lowest 5-year survival rates, i.e., Only 9% of people with PDAC are able to survive for five years after diagnosis. Pancreatic cancer was accountable for nearly 500,000 deaths globally in 2020, as reported by GLOBOCAN 2020. PDAC is highly heterogeneous, leading to differences in oncogenesis and varying survival rates among patients. Consequently, this heterogeneity poses significant clinical challenges, including inaccurate diagnoses and inadequate treatment approaches. [2]. Contemporary medical and molecular diagnostic techniques, such as computed tomography (CT), magnetic resonance imaging (MRI), and endoscopic ultrasound (EUS) accompanied by fine-needle aspiration (FNA), are currently available. However, these techniques *offer limited information regarding tumor aggressiveness and the probable disease* prognosis, posing a significant challenge to the development of an accurate treatment regimen[3]. As a result, the prognosis post-surgery remains, in most cases, uncertain[4] (Guillén-Ponce et al., 2017). A case in point, is the CA 19-9, as discussed in this study. Which is a widely used cancer marker for monitoring treatment Responses [5]. And This informs the treatment strategy, however, this biomarker has demonstrated high false positive and high false negative results[4].

### *1.1 Multi-Omics*

The advancement in high-throughput technologies[6], and explosive growth in biological data collection, e.g., Several thousand biological samples have been profiled and made publicly available by The Cancer Genome Atlas (TCGA) and The International Cancer Genome Consortium (ICGC) have allowed researchers to understand the molecular bases of the genetic disorders, facilitating effective and personalized diagnosis and treatments in case of many cancers, including PC[7][8][9][10]. However, cancer research that concentrates on only one aspect of biological data (single-omics) has only furnished limited insights into the causes of cancer development and the advancement of tumors[11]. These single omics studies have often resulted in different and at times conflicting patient classifications[12]. And as such many multi-omics studies of various cancers have facilitated a deeper understanding of genomics, epigenomics, and transcriptomics deregulations in malignancies[13]. One such method is the use of the machine learning model called Autoencoders, which is an Artificial Intelligence method of relearning the latent space or manifold of the high-dimensional space to extract meaningful information from large data bodies. The omics data are charlatanically very complex in their dimensions and hence application of AE In such situations has been extensively studied.

*Aadil Rashid Bhat
*E-mail address: aadil.bhat@uok.edu.in*

### *1.2 Dimensionality Reduction and Machine Learning*

Dimensionality reduction is a widely adopted technique in data science and machine learning that involves the process of converting information from a space with a large number of dimensions into a smaller subspace with fewer dimensions. The primary objective of this technique is to preserve the crucial attributes of the original data while reducing its complexity and computational requirements One way to accomplish this is by mapping the important features of the data onto a subspace with fewer dimensions that captures the essential characteristics of the original data that are relevant to a given use case or analysis.[14]. Dimension reduction techniques have been used for classification, visualization, and data compression in many fields including bioinformatics[15]. Well-known techniques for data analysis such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Multi-Dimensional Scaling (MDS)[14] have been extensively studied. However, these methods suffer from different shortcomings like the inability to capture global structures[16], information loss, and processing of large-scale datasets[17]. Apart from DM methods, many methods have been devised for the integration of multi-omics data these methods attempt to find the complementary signals from different omics functional levels to better understand the molecular characterization of underlying conditions like cancer. This combined analysis has the potential to reveal novel biomarkers for better characterization and homogenous classification of the cancer types and subtypes and an assistant in personalized care and better treatment plans. multi-omics for cancer patient stratification has also been widely studied. For example, The intNMF[18], which is a non-negative matrix factorization method, which does not assume any distributional form of data, The LRACluster[19] a probabilistic integrative mode based on low-rank approximation, which assumes and estimates the principal latent subspace for the entire data, The Mixkernal[20], which computes similarity matrices from kernels which are then combined to obtain a combined Similarity matrix. The SNF, which is a popular similarity network fusion method that uses graphs to model patient-patient similarity using multi-omics data, and the RCGAA[21] a generalized canonical correlation analysis framework that allows for choosing various parameters like scheme functions and shrinkage constants. Even though unsupervised subtyping revealed the molecular diversity among PDAC patients, the survival outcomes varied widely within each subtype. As a result, there is no significant difference in prognosis among the subtypes identified using these methods [22]. More recently Artificial intelligence-based methods which have shown remarkable performance in other data extensive fields have been applied to the problem of dimensionality reduction and data integration. In this study, we used Deep hierarchical Variational Autoencoders (DVAE) for dimensionality reduction and multi-omics integration to learn homogenous subtypes of pancreatic cancer.

## 2. Description of model for PC classification

### *2.1 Proposed model.*

Our proposed model is a multi-level deep Variational Autoencoder (VAE) based integration and dimensionality reduction framework. The standard VAE[23] is a probabilistic deep learning method that is used to extract the low-dimensional data manifold from high-dimensional datasets. Instead of representing each input $x_i$ as a singular value, the VAE encodes it as a distribution characterized by its mean and standard deviation across a latent space. VAEs are typically composed of two networks an input network that encodes the data, and a decoder counterpart that aims to reproduce the original input data from the low-dimensional embeddings, see Fig 1.
The VAE approximates the latent distribution by minimizing the kl divergence score, as in Eq. (1).

$$D_{KL}(q_\phi(z) \parallel p_\theta(z)) \tag{1}$$

Where $q_\phi(z)$, called a variational distribution, is the estimation of the true but intractable posterior $p_\theta(z)$. Recently, Maximum Mean Discrepancy (MMD)[24] function in comparison to KL divergence was shown to produce better results for approximating the true posterior. MMD, which is given by Eq. (2),

$$MMD(p(z)||q(z)) = \mathbb{E}_{(p(z),p(z'))}[k(z,z')] + \mathbb{E}_{(q(z),q(z'))}[k(z,z')] - 2\mathbb{E}_{(p(z),q(z'))}[k(z,z')] \tag{2}$$

states that two distributions can only be considered identical if their moments are equal. Therefore, we can measure divergence by comparing the moments of two distributions $p(z)$ and $q(z)$. By using kernel embedding, MMD can efficiently accomplish this

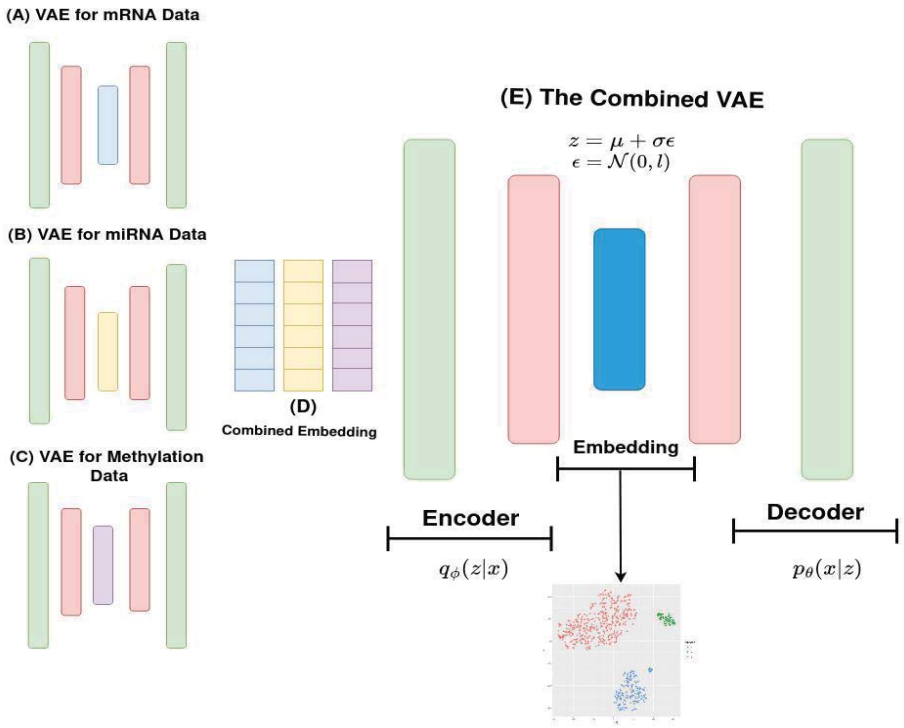


Fig. 1. The Architecture of Deep Hierarchical Variational Autoencoder for PAAD. A separate VAE is built for (A) mRNA (B) microRNA (C) Methylation. The individual omics embeddings are then combined (E) to train the combined VAE model (E).

In this study, we trained three different MMD-based VAEs for each omics data and the outputs of which were fed to another VAE to generate the final embeddings which capture the latent manifold of the entire data and are subsequently employed for finding homogenous subtypes of the PC.

### *1.3 Datasets*

In the present investigation, we employed a comprehensive approach for pancreatic cancer (PC) subtyping utilizing four distinct omics data sets. Specifically, we retrieved data about mRNA expression, microRNA expression, DNA methylation array, as well as clinical parameters of a cohort of 150 patients who underwent surgical removal of their primary pancreatic ductal adenocarcinoma (PDAC from The Cancer Genome Atlas (TCGA) Pancreatic Adenocarcinoma (PAAD) database [7].

Training Dataset: We obtained transcriptome-wide information on 177 patients by downloading data from The Cancer Genome Atlas (TCGA) Pancreatic Adenocarcinoma (PAAD) cohort using the R package TCGA-Assembler[25]. The information set consisted of three types of biological data: mRNA sequencing (mRNA-Seq), microRNA sequencing (microRNA), and DNA methylation array data. The Illumina HiSeq platform was utilized to generate both mRNA-Seq and microRNA data, while the Illumina Infinium HumanMethylation450 BeadChip platform was used to obtain the DNA methylation data. The mRNA-Seq data, as per TCGA, was processed and normalized by Expectation Maximization (RSEM) [26]. Likewise, the RPM normalization method was applied to the microRNA-Seq data. Only patient samples that had complete data for all three types of multi-omics and clinical information (146 samples) were kept. Any genes that had missing data greater than 20% for DNA methylation, as well as any genes from mRNA and microRNAs that had zero values greater than 20% among the retained samples, were excluded. [27][20] DNA methylation genes exhibiting a proportion of missing values less than or equal to 20% were subjected to imputation using the R package impute. Subsequently, to standardize the values of the mRNA and microRNA datasets, a log transformation was utilized. This data preprocessing strategy was implemented to reduce the impact of missing data on downstream analyses, as well as to facilitate meaningful comparisons between gene expression profiles across multiple samples.

Test Datasets: To corroborate our findings and evaluate the classification efficacy of our model, we procured two external datasets from the Gene Expression Omnibus (GEO), with the accession numbers GSE62498 and GSE62452, respectively.[28]. The GEO GSE62452 mRNA microarray dataset was procured from the Affymetrix GeneChip platform and was subjected to the robust multi-array average (RMA) normalization technique. Subsequently, the data generators obtained the average expression

values of each gene from the multiple corresponding probe sets. The expression values were then subjected to a logarithmic transformation. The GEO GSE62498 microRNA dataset, on the other hand, was obtained from the Nanostring nCounter Platform. The dataset underwent normalization through the geometric mean, followed by a logarithmic transformation using the formula $\log_2(x+1)$.

*2.2 Model Training*

Four omics pre-processed TCGA PC data for a total of 146 patients were used as input to the Autoencoders. The DVAE architecture was constructed utilizing the Keras library in the Python programming language. The design consists of a symmetric encoder and decoder neural network architecture with two hidden layers and a maximum mean discrepancy (MMD) loss function. ReLU, a commonly used nonlinear function, was employed as the activation function in each layer except for the output layer, which employed the sigmoid function for each layer the output y, given the input $x$, is calculated as in Eq. (3).

$$y_i = relu\left(\dot{W}_i x + b_i\right) = f_i(x) \tag{3}$$

So, the output $\hat{x}$, is given by Eq. (4).

$$\hat{x} = sigmoid\left(f_3\left(f_2\left(f_1(x)\right) + b\right)\right) = F(x) \tag{4}$$

After the model reduced the number of features to 500, a univariate Cox-PH model was generated on each feature, and the top features associated with survival (P-values * 0.05) were identified using the R survival package. K-means was next used to cluster the data into two different survival subgroups.

## 3. Results

Subsequently, we retrieved the omics characteristics with the most unique expression patterns between the two subtypes (i.e., high-risk or low-risk) that had been previously identified. The ANOVA-F test was applied to assess the features that were significantly related to each subgroup, and the selection of features was based on the average precision obtained from 5-fold cross-validation on the training set. These features were then utilized to train a random-forest classifier. Furthermore, the effectiveness of our approach was confirmed by evaluating two external datasets. The Kaplan–Meier plot of the training dataset and the two external datasets are presented below in Fig 2

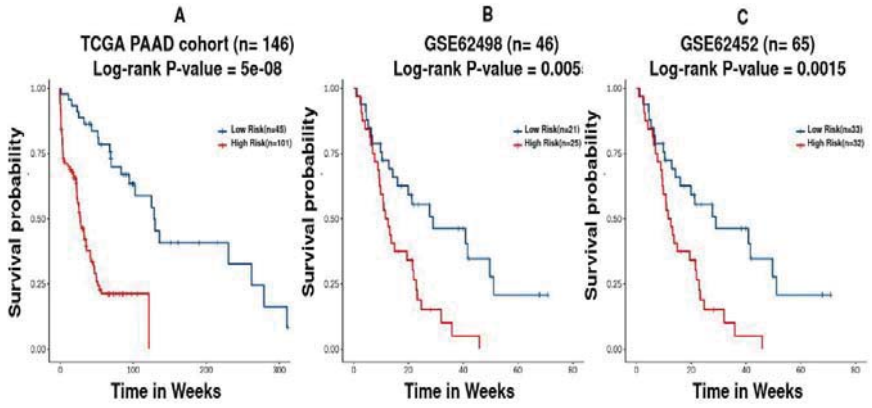


Fig. 2. The Kaplan-Meier survival plots of the subgroups for the (A) TCGA-PAAD cohort and two external datasets (B) GSE62498 with 46 patients and p-value of 0.0055 (C) GSE62452 with 65 patients and p-value of 0.0015.

*3.1 Comparison with other methods*

We then compared the performance of our model against 6 other integration methods based on different statistical frameworks like Matrix factorization, consensus clustering and co-inertia analysis, and similarity matrix.

we assessed the effectiveness of integration and dimensionality reduction algorithms using the log-rank p-value and concordance index as the evaluation metrics. see the comparison table and plots in Fig 3. Our proposed model demonstrated superior performance to other existing multi-omics models, as evidenced by a log-rank p-value of 5e-08 and a c-index of 0.6505, as depicted in Fig 3. These findings suggest that our model has the ability to extract meaningful subspace manifold and possesses discriminative power to accurately classify new, unseen single omics datasets into the identified subtypes.

**(A) The Comparison of various Intgeration methods**

| | P-log Value | Concordance index |
|---|---|---|
| Our Method | 5e-08 | 0.6505 |
| MCIA | 0.008 | 0.5606 |
| SNF | 1e-05 | 0.6131 |
| RGCAA | 3e-06 | 0.6374 |
| intNMF | 0.0002 | 0.5898 |
| MixedKernal | 6e-05 | 0.6175 |
| LRACluster | 0.009 | 0.5664 |

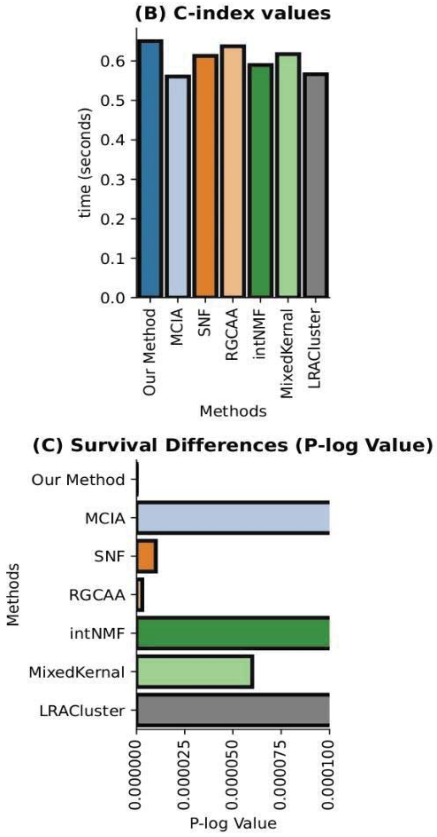


Fig. 3. The performance comparison of our model against six other related methods (a) table summarizing the findings, and the bar plots of the (b) log-Rank P-values and (c) concordance index.

## 4. Discussion and Conclusion

Over the past few years, there has been significant research dedicated to the detection of molecular and biomarker features of cancers, which includes pancreatic cancer (PC). A study conducted in 2012 on 1027 cases of PC and 1031 controls of Han Chinese patients identified a noteworthy correlation between the incidence of pancreatic cancer and the copy number of CNVR2966.1 located at 6q13.[29]. Another study in 2015 established a link between amplification in the MYC gene and shorter patient survival duration.[30] Multi-omics studies have been employed to improve diagnostic tools, such as Comp Cyst, which is a machine learning-based test used to manage patients with Pancreatic cystic lesions (PCLs) and is estimated to help avoid 60% of unnecessary surgeries.[31] . Multi-omics has also been used to study cancer patient stratification. For example, the iCluster method utilized gene expression and copy number variation information to identify different subtypes of breast and lung cancer. This approach demonstrated that utilizing multiple sources of information (multi-omics) results in more informative subtypes compared to using only one source (single omics). The Similarity Networks Fusion (SNF) technique was also applied to identify molecular subtypes of pancreatic cancer using a combination of proteins, mRNAs, DNA methylation, and microRNA profiles[8]. Although unsupervised subtyping helped using non-AI methods to identify molecular diversity in PDAC patients, the patients in each subtype still exhibited a wide range of survival outcomes, and the disparities among subtypes were not statistically significant.[22].

In this study, we used the ability of the Deep Variational Autoencoder to extract the latent clustering patterns within the Pancreatic Cancers patients to identify and characterize the two prognosis subtypes. These features were then used to train a machine learning classification model to classify the unseen Pancreatic cancer datasets into the identified subtypes.

## References


1. M. D. Siegelin and A. C. Borczuk, "Epidermal growth factor receptor mutations in lung adenocarcinoma," *Lab. Investig.*, vol. 94, no. 2, pp. 129–137, 2014, doi: 10.1038/labinvest.2013.147.
2. D. P. Ryan, T. S. Hong, and N. Bardeesy, "Pancreatic Adenocarcinoma," pp. 1039–1049, 2014, doi: 10.1056/NEJMra1404198.
3. S. S. Vege, B. Ziring, R. Jain, and P. Moayyedi, "American gastroenterological association institute guideline on the diagnosis and management of asymptomatic neoplastic pancreatic cysts," *Gastroenterology*, vol. 148, no. 4, pp. 819–822, 2015, doi: 10.1053/j.gastro.2015.01.015.
4. J. Bla, "Diagnosis and staging of pancreatic ductal adenocarcinoma," pp. 1205–1216, 2017, doi: 10.1007/s12094-017-1681-7.
5. V. Eijck, A. P. Stubbs, and Y. Li, "Jo ur na l P re of," *ISCIENCE*, p. 103415, 2021, doi: 10.1016/j.isci.2021.103415.
6. [Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biol.*, vol. 18, no. 1, pp. 1–15, 2017, doi: 10.1186/s13059-017-1215-1.
7. B. J. Raphael *et al.*, "Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma," *Cancer Cell*, vol. 32, no. 2, pp. 185-203.e13, 2017, doi: 10.1016/j.ccell.2017.07.007.
8. M. Sinkala, N. Mulder, and D. Martin, "Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics," pp. 1–14, 2020, doi: 10.1038/s41598-020-58290-2.
9. T. Golan and M. Javle, "DNA Repair Dysfunction in Pancreatic Cancer : A Clinically Relevant Subtype for Drug Development," vol. 15, no. 8, pp. 1063–1069, 2017, doi: 10.6004/jnccn.2017.0133.
10. T. J. Grant, K. Hua, and A. Singh, *Molecular Pathogenesis of Pancreatic Cancer*, 1st ed. Elsevier Inc., 2016.
11. M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions," *Nat. Rev. Genet.*, vol. 16, no. 2, pp. 85–97, 2015, doi: 10.1038/nrg3868.
12. B. Wang *et al.*, "Similarity network fusion for aggregating data types on a genomic scale," *Nat. Methods*, vol. 11, no. 3, pp. 333–337, 2014, doi: 10.1038/nmeth.2810.
13. Y. Hasin *et al.*, "A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data," *Genome Biol.*, vol. 19, no. 1, pp. 71–86, 2018, doi: 10.1093/biostatistics/kxx017.
14. [R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 56–70, 2020, doi: 10.38094/jastt1224.
15. [E. Postma and E. Postma, "Dimensionality Reduction : A Comparative Review Dimensionality Reduction : A Comparative Review," 2009.
16. [J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science (80-. ).*, vol. 290, no. 5500, pp. 2319–2323, 2000, doi: 10.1126/science.290.5500.2319.
17. E. Becht *et al.*, "A n a ly s i s Dimensionality reduction for visualizing single-cell data using UMAP," vol. 37, no. 1, 2019, doi: 10.1038/nbt.4314.
18. P. Chalise and B. L. Fridley, "Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm," *PLoS One*, vol. 12, no. 5, pp. 1–18, 2017, doi: 10.1371/journal.pone.0176278.
19. D. Wu, D. Wang, M. Q. Zhang, and J. Gu, "Fast dimension reduction and integrative clustering of multi-omics data using lowrank approximation: Application to cancer molecular classification," *BMC Genomics*, vol. 16, no. 1, pp. 1–10, 2015, doi: 10.1186/s12864-015-2223-8.
20. B. Zhu *et al.*, "Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers," *Sci. Rep.*, vol. 7, no. 1, pp. 1–13, 2017, doi: 10.1038/s41598-017-17031-8.
21. M. Tenenhaus, A. Tenenhaus, and P. J. F. Groenen, "Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods," *Psychometrika*, vol. 82, no. 3, pp. 737–777, 2017, doi: 10.1007/s11336-017-9573-x.
22. A. J. Aguirre, "Refining Classification of Pancreatic Cancer Subtypes to Improve Clinical Care," *Gastroenterology*, vol. 155, no. 6, pp. 1689–1691, 2018, doi: 10.1053/j.gastro.2018.11.004.
23. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes." 2014.
24. A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," *Adv. Neural Inf. Process. Syst.*, vol. 1, pp. 513–520, 2007, doi: 10.7551/mitpress/7503.003.0069.
25. L. Wei, Z. Jin, S. Yang, Y. Xu, Y. Zhu, and Y. Ji, "TCGA-assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data," *Bioinformatics*, vol. 34, no. 9, pp. 1615–1617, 2018, doi: 10.1093/bioinformatics/btx812.
26. B. Li and C. N. Dewey, "RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome," *Bioinforma. Impact Accurate Quantif. Proteomic Genet. Anal. Res.*, pp. 41–74, 2014, doi: 10.1201/b16589.
27. S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in multi-omics data integration methods," *Front.*

*Genet.*, vol. 8, no. JUN, pp. 1–12, 2017, doi: 10.3389/fgene.2017.00084.

28. J. Luo *et al.*, "Prognostic and predictive value of the novel classification of lung adenocarcinoma in patients with stage IB," *J. Cancer Res. Clin. Oncol.*, vol. 142, no. 9, pp. 2031–2040, 2016, doi: 10.1007/s00432-016-2192-6.
29. L. Huang *et al.*, "Copy number variation at 6q13 functions as a long-range regulator and is associated with pancreatic cancer risk," *Carcinogenesis*, vol. 33, no. 1, pp. 94–100, 2012, doi: 10.1093/carcin/bgr228.
30. A. K. Witkiewicz *et al.*, "Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets," *Nat. Commun.*, vol. 6, pp. 1–11, 2015, doi: 10.1038/ncomms7744.
31. E. J. Hoorn, "Multicenter Paper," *Physiol. Behav.*, vol. 176, no. 1, pp. 100–106, 2017, doi: 10.1126/scitranslmed.aav4772.A.