BMC Bioinformatics

RESEARCH Open Access

Integrating genetic and gene expression data in network-based stratification analysis of cancers

Kenny Liou[2] and Ji-Ping Wang[1*]

*Correspondence:
jzwang@northwestern.edu

[1] Department of Statistics and Data Science, Northwestern University, 2006 Sheridan Road, Evanston, IL 60208, USA
[2] Stevens Neuroimaging and Informatics Institute, University of Southern California, 2025 Zonal Ave, Los Angeles, CA 90033, USA

Abstract

Background: Cancers are complex diseases that have heterogeneous genetic drivers and varying clinical outcomes. A critical area of cancer research is organizing patient cohorts into subtypes and associating subtypes with clinical and biological outcomes for more effective prognosis and treatment. Large-scale studies have collected a plethora of omics data across multiple tumor types, providing an extensive dataset for stratifying patient cohorts. Network-based stratification (NBS) approaches have been presented to classify cancer tumors using somatic mutation data. A challenge in cancer stratification is integrating omics data to yield clinically meaningful subtypes. In this study, we investigate a novel approach to the NBS framework by integrating somatic mutation data with RNA sequencing data and investigating the effectiveness of integrated NBS on three cancers: ovarian, bladder, and uterine cancer.

Results: We show that integrated NBS subtypes are more significantly associated with overall survival or histology. Specifically, we observe that integrated NBS subtypes for ovarian and bladder cancer were more significantly associated with patient survival than single-data type NBS subtypes, even when accounting for covariates. In addition, we show that integrated NBS subtypes for bladder and uterine are more significantly associated with tumor histology than single-data type NBS subtypes. Integrated NBS networks also reveal highly influential genes that span across multiple integrated NBS subtypes and subtype-specific genes. Pathway enrichment analysis of integrated NBS subtypes reveal overarching biological differences between subtypes. These genes and pathways are involved in a heterogeneous set of cell functions, including ubiquitin homeostasis, p53 regulation, cytokine and chemokine signaling, and cell proliferation, emphasizing the importance of identifying not only cancer-specific gene drivers but also subtype-specific tumor drivers.

Conclusions: Our study highlights the significance of integrating multi-omics data within the NBS framework to enhance cancer subtyping, specifically its utility in offering profound implications for personalized prognosis and treatment strategies. These insights contribute to the ongoing advancement of computational subtyping methods to uncover more targeted and effective therapeutic treatments while facilitating the discovery of cancer driver genes.

Keywords: Cancer subtyping, Network-based stratification, Omics, Survival analysis, Data integration, Precision medicine

Introduction

Cancer, a multifaceted disease, exhibits remarkable genetic diversity among patients, rendering the development of effective treatments a formidable challenge. To gain valuable insights into this complexity, large-scale initiatives such as the International Cancer Genome Consortium (ICGC) [1] and The Cancer Genome Atlas (TCGA) [2] meticulously profile genomic, transcriptomic and epigenomic data. This includes data on DNA methylation, microRNA expression, and protein expression, unveiling a treasure trove of information. Both initiatives are driving research in pan-cancer analysis, biomarker identification, and computational modeling, fueled by the sheer volume of available data [3–6]. Consequently, the massive amount of data from these projects has spurred a demand for informatics solutions to unearth molecular pathways dictating tumor progression, understanding disease populations, and designing precision-targeted treatment strategies [7].

In cancer informatics, a pivotal objective is to categorize or stratify heterogeneous cancer tumors into clinically and biologically meaningful subtypes using molecular profiling data. Historically, this endeavor primarily leveraged mRNA expression data, yielding valuable insights into cancers like ovarian cancer [8] and breast cancer [9]. However, it has fallen short in other cancer types, such as colorectal cancer, where the association between molecular subtypes and clinical phenotypes remains elusive [10]. This limitation may be attributed to issues like sample quality and overfitting, inherent to gene expression analysis [11]. Data collection sequencing protocols in cancer initiatives such as TCGA can introduce technical and biological biases, such as elevated noise and tumor heterogeneity, complicating the interpretation of the results and hindering the identification of high-resolution cancer subtypes [12, 13].

As the landscape of genome-scale data diversifies and expands, there emerges an urgent need to develop methodologies that effectively integrate multi-omics data to enhance tumor subtyping, predict prognosis, and identify relevant biomarkers. [14, 15]. Multi-omic research in cancer genetics offers promising results, such as in breast cancer [16, 17]. Typically, multi-omic integration methods synthesize data after processing -omics data types separately or using different -omics data types for different parts of the methodology pipeline [14]. The advantage of multi-omics is that it considers the potential interactions between various molecular layers, providing a holistic perspective that can yield different results compared to single-data-type methods [18]. By capturing the interdependencies across multiple molecular layers, multi-omics offers a powerful tool to account for the influence of biological disease pathways on different omic data types [19].

Recent years have witnessed the widespread adoption of gene interaction networks, with cancer being increasingly recognized as a network-driven disease [20]. Somatic mutation profiles harbor cancer-driver genes capable of instigating genetic mutations in other genes. Network-based Stratification (NBS) [21] is an approach that melds gene interaction networks with somatic mutation profiles. This fusion involves mapping somatic mutation profiles onto a cancer network and propagating these mutations

throughout the network to create smoothed network profiles. NBS uses these profiles to stratify tumors by clustering patients with similar smoothed network profiles.

In this study, we propose a novel multi-omics methodology grounded in NBS for tumor stratification, fusing somatic mutation profiles with RNA gene expression profiles in a manner that, to our knowledge, has not been explored in the current NBS literature. We fuse somatic mutation and gene expression profiles before applying network propagation to the integrated profiles to explore if integrating profiles before network smoothing can generate informative cancer subtypes. By integrating these two data types before network smoothing, we have successfully generated robust, biologically informative, and clinically significant tumor subtype clusters. Our methodology was tested using the TCGA ovarian, uterine, and bladder carcinoma cohorts, yielding compelling results that can shed new light on understanding cancer population structures through an integrated approach.

Methods

Integrating genetic profiles

Ovarian serous adenocarcinoma, uterine endometrial carcinoma, and bladder urothelial carcinoma somatic mutation and gene expression data were downloaded from the TCGA Genomic Data Commons Data Portal. https://portal.gdc.cancer.gov/. Accessed 10 June 2023. Only individuals with somatic mutation and gene expression data were retained for the experiment. 279 ovarian cancer, 318 uterine endometrial carcinoma, and 399 bladder urothelial carcinoma patients from the TCGA cohorts were used in subsequent analysis. Somatic mutation profiles are binary vectors where a '0' or '1' for whether a gene for that individual has no mutation or a mutation, respectively. Gene expression profiles are continuous TPM normalized values where the value represents the level of gene expression. We were inspired by advancements in machine learning that integrate binary and continuous data types to improve diagnostic accuracy and cancer detection through linear models [22–24]. The gene expression profiles were min-max normalized gene by gene to match the 0 to 1 range of somatic profiles. Somatic mutation and gene expression profile integration can be represented through this formula:

$$S_i = \beta \times p_i + (1 - \beta) \times q_i, \tag{1}$$

where $0 < \beta < 1$ is a tuned hyperparameter chosen by the user to linearly combine the somatic mutation profile p_i and the normalized gene expression profile q_i to result in the integrated profile S_i for individual i. After performing a β hyperparameter selection procedure for each of the ovarian cancer, bladder cancer, and uterine cancer cohorts, we utilized tuned β values of 0.8, 0.3, and 0.1, respectively. The value of β for ovarian and bladder cancers were chosen to give the most consistently significant p-values across all cluster numbers (K) in the log-rank test from Kaplan Meier survival analysis or the log-likelihood ratio test from Cox regression survival analysis. Survival analysis for uterine cancer was impeded by the notably low mortality rate within the cohort, alongside the absence of significant associations with survival observed among both single-data type and integrated subtypes. To assess the association of uterine cancer subtypes, we examined the association of TCGA subtypes to our generated Multi-NBS subtypes (χ^2

association test statistic) to derive the value of β. Additional details about TCGA are provided at https://www.cancer.gov/ccg/research/genome-sequencing/tcga [2].

Gene interaction network

The constructed network is derived from PCNet [25], a network with 19,781 genes and 2,724,724 interactions that is then filtered for cancer-specific genes and interactions in at least one of these four sources: [26–29]. The resulting cancer subnetwork contained 2291 nodes. This network was used for all three methods of NBS. Further details on constructing this network can be found in a previous NBS publication [30].

Network propagation

The integrated profiles are mapped onto the gene interaction network and then network propagation [31] is applied to diffuse the signals across the network. Let m be the number of genes, n be the number of patients, F_0 be the initial $n \times m$ (patient $\times$ gene matrix), and A be the symmetric adjacency matrix ($m \times m$) representing the gene-gene interaction network obtained above. Network propagation follows the following iterative procedure:

$$F_{t+1} = \alpha F_t A + (1 - \alpha) F_0, \tag{2}$$

where $\alpha = 0.7$, which is derived from benchmarking results reported in earlier NBS publications [21]. The network is propagated till F_t converges ($|F_{t+1} - F_t < 0.001|$). After convergence, the resulting matrix F_t was quantile normalized by row (patient) to ensure each patient followed the same distribution. F represents the final normalized and smoothed integrated matrix.

Network-regularized NMF

Non-negative matrix factorization (NMF) decomposes a matrix into two non-negative matrices whose product results in the original matrix. Network-regularized NMF is an extension of NMF that constrains NMF to respect the network structure [32–34]. The following objective function is minimized:

$$min_{W,H>0}\left\{\left\|F - WH\right\|^2 + trace(W^t J W)\right\}. \tag{3}$$

F is approximately decomposed into the product of non-negative matrices W (m by K matrix) and H (K by n matrix). W is a collection of basis vectors ("meta-genes") and H represents the basis loadings. K controls the dimension reduction and we used values of $K = 2, 3, 4, 5, 6, 7, 8$ in the following context. The $trace(W^t J W)$ function is responsible for constraining the basis vectors in W to respect the local neighboring network structure. J represents the graph Laplacian of the k-nearest neighbor network and we used $k = 11$ as described in previous work for consistency [21, 35, 36].

Consensus clustering

We used consensus clustering [37] to ensure robust clustering to achieve the final patient cluster assignments. We performed network-regulated NMF using a random sampling without replacement of 80% of patients. This was repeated 100 times as described

previously [21]. The collection of 100 clustering results was used to construct a similarity matrix that recorded the frequency with which patient pairs had the same cluster assignment from all iterations where both patients in the pair were sampled.

Implementation of integrated network-based stratification

The implementation of integrated network-based stratification is a Python 3.10 version based on an existing Python 2.7 implementation of NBS [30]. We provide an implementation of integrated network-based stratification at the Multi-NBS repository.

Cluster analysis

We used Silhouette scores [38] as internal cluster measures. The Silhouette score calculates how well a patient is affiliated with its cluster compared to neighboring clusters. We also used Adjusted Mutual Information (AMI) [39] to assess cluster similarity between clusters formed through different data-type generated clusters.

Survival analysis

Survival analysis was performed through the lifelines [40] and scikit-survival package [41]. Kaplan-Meier survival curves [42] were fitted to the subtypes generated and log-rank tests were performed on single-data type and integrated NBS subtypes to assess the association between the subtypes and survival. We also fitted a semi-parametric Cox proportional hazard model [43]. The Cox hazard model can be represented with a hazard function $h(t|X_i)$ at time t for an individual i, given p covariates, denoted by X_i:

$$h(t|X_i) = h_0(t) \exp \sum_{j=1}^{p} \beta_j X_{ij}. \tag{4}$$

We included covariates such as age, race, and gender in addition to the cluster assignments to assess cluster assignment influence on survival. The log-likelihood ratio test compares the full model with subtype assignments and clinical covariates against a null model. The log-likelihood ratio test and the associated p-value provide an estimate of the predictive power of the full model (clinical covariates and subtype assignments) compared to the null model. The concordance index assesses the discriminatory power of the model by evaluating the ability of the model to correctly rank the survival time of pairs of patients. Maximizing the concordance index indicates the model is discriminative of early events (which are associated with higher-risk patients) and later events.

Association with TCGA subtypes

Association with TCGA subtypes is calculated through Pearson's χ^2 test of independence between computed integrated subtypes or single-data type subtypes and documented TCGA subtypes. The documented TCGA subtypes are obtained through the R "TCGAbiolinks" library [44–46], which are determined through genomic, transcriptomic, and proteomic characterization of tumors using array and sequence-based technologies by The Cancer Genome Atlas Research Network [47–49].

Identifying high-scoring genes

The integrated profiles were first propagated through the network propagation process described above. The propagated integrated profiles are then grouped by the subtype assignments generated using Multi-NBS. After being grouped by subtype the propagated integrated profiles are averaged by number of patients in each subtype to find genes with the highest average hybrid score in the network for each integrated profile subtype. We refer to the score as a "hybrid score" because the network values for integrated profile networks are a combination of gene expression and somatic mutation data.

For somatic mutation high-scoring genes, the exact same process is applied except we only used somatic mutation profiles and the subtype assignments generated from using somatic mutation profile NBS. We first applied propagation to the somatic mutation profiles. The propagated somatic mutation profiles are grouped by the subtype assignments given by NBS that used only somatic mutation profiles. After being grouped, the propagated somatic mutation profiles are averaged by number of patients in each subtype to find genes with the highest average mutation score in the network for each somatic mutation profile subtype. Note that we refer to the somatic mutation scores as mutation scores because the network values are derived from solely somatic mutation data.

Pathway enrichment analysis

We performed a gene set enrichment analysis (GSEA) using smoothed integrated subject profiles to identify pathways enriched in specific subtypes. GSEA evaluates whether a predefined gene set is significantly enriched at the top or bottom of a ranked gene list using a modified Kolmogorov-Smirnov (KS) statistic. It calculates an enrichment score by comparing the observed distribution of gene ranks to a null distribution generated through permutations, with false discovery rate (FDR) correction applied to assess significance. We applied GSEA using the "clusterProfiler" package in R [50, 51] and gene sets for pathway analysis were obtained from the Kyoto encyclopedia of genes and genomes (KEGG) pathway database [52]. FDR-adjusted p-values were reported to identify statistically significant pathways, with a threshold set at $FDR < 0.05$.

Results

For the three cancer types under consideration, ovarian, uterine, and bladder carcinoma, only individuals extracted from the TCGA database with both somatic mutation and gene expression profiles were retained, giving 279 ovarian carcinoma, 318 uterine endometrial carcinoma, and 399 bladder urothelial carcinoma patients. We did not remove individuals with fewer than 10 somatic mutations. Multi-omic Network-Based Stratification (Multi-NBS) merges somatic mutation profiles with RNA sequencing (RNA-seq) gene expression data through a linear combination to creating an integrated profile (Fig. 1A). The integrated profile is subsequently projected to a gene interaction network and the influence of mutated genes is propagated to its neighboring nodes through network propagation (Fig. 1B). The underlying gene interaction network used in this context is a sub-network comprised of cancer-related genes, which was constructed from PCNet [25], but filtered using four databases containing cancer-specific genes

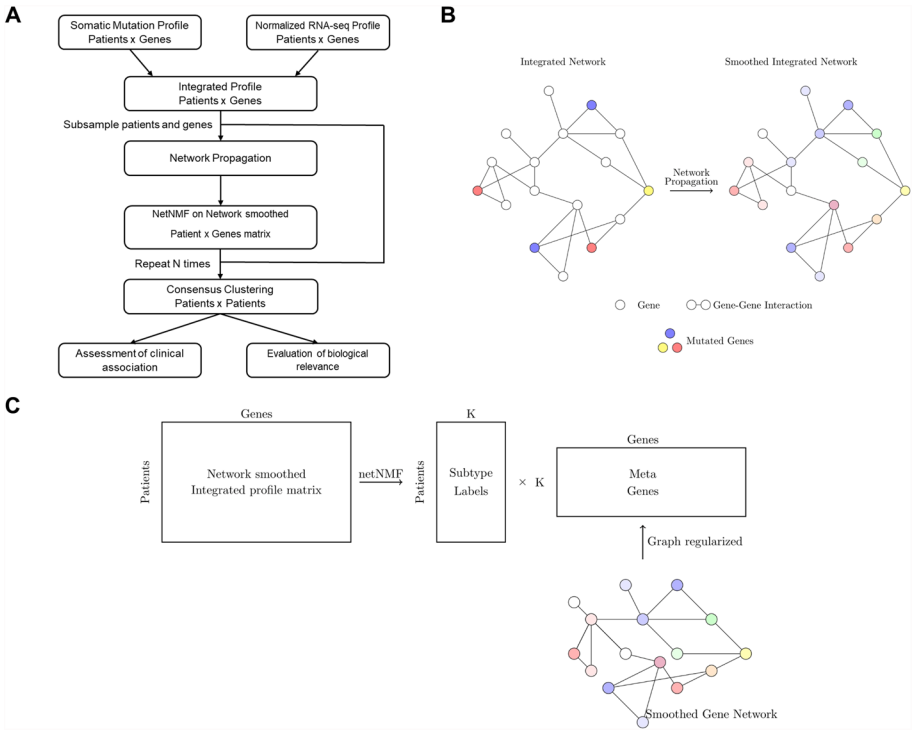

Fig. 1 Integrated NBS workflow. **A** A workflow of an integrated approach to network-based stratification. **B** A schematic of network propagation on a gene interaction network. **C** A representation of non-negative matrix factorization constrained by the gene interaction network structure

and interactions [26–29]. This "smoothed" integrated network is decomposed through a variant of non-negative matrix factorization that respects the network structure to derive subtype assignments (Fig. 1C). We use consensus clustering to refine subtype assignments by aggregating results from repeated subsampling, yielding a single robust stratification result. To assess the efficacy of the subtypes generated by Multi-NBS, we conduct a rigorous performance evaluation, comparing them against subtypes derived from single-data type NBS. Additionally, we employ benchmarking procedures to identify the optimal β values for integration, ensuring the robustness and informativeness of our results (See Methods).

In the following we assess the efficacy of Multi-NBS clusters in terms of clustering metrics, clinical significance in predicting survival time, and association with histological characteristics by benchmarking with clusters obtained from single data type. In addition we will investigate whether the identified Multi-NBS clusters can reveal new high-scoring genes within the networks to that have cancer relevance.

Cluster evaluation

We initially assessed the comparability of clusters generated by Multi-NBS to those from single-data-type approaches using Silhouette scores [38], a widely utilized clustering metric across various disciplines, to determine the suitability of Multi-NBS clusters for further clinically relevant downstream analyses. Higher Silhouette scores (values closer to 1) indicate well-formed clusters while lower scores (values closer to -1) indicate

poorly-formed clusters. Values close to 0 in Silhouette scores are still generally considered acceptable for classification. We plotted the Silhouette scores as a function of the number of clusters generated under each method (Fig. 2). Somatic mutation clusters appeared to have the poorest formed clusters while RNA-seq and Multi-NBS clusters appeared to generate more well-formed clusters overall (Fig. 2A, B). RNA-seq seemed to generate the most well-formed clusters (Fig. 2C). The Multi-NBS clusters did not show a uniformly better pattern than the other two. Instead, in ovarian and bladder cancers, we see that the Multi-NBS Silhouette score curves were similar to RNA-seq clusters, while in uterine cancer, the Multi-NBS Silhouette score was observed to be between the two single-data-type Silhouette scores. Although Multi-NBS did not consistently outperform other methods, clustering metrics like Silhouette scores are broadly applicable and may not offer definitive validation on their own. Nonetheless, this analysis provides a solid preliminary assessment, leading us to conclude that clusters derived from integrated profiles are sufficiently well-structured to justify further, more tailored analyses to explore their biological and clinical relevance.

We next compared the similarity of cluster contents using Adjusted Mutual Information (AMI) [39]. AMI is a measure of cluster independence, where a score of 0 means the clusters are completely independent and a score of 1 means the clusters are completely identical. Across all cancer types, somatic mutation and RNA-seq clusters showed minimal overlap (Fig. 3). The Multi-NBS and RNA-seq clusters showed relatively higher cluster similarity and the AMI values are 0.28, 0.82 and 0.8 respectively for ovarian, bladder and uterine cases, reflecting the impact of RNA-seq data in the calculation of blended gene profile S_i (see Eq. 1) (i.e., $1 - \beta$ are 0.2, 0.7 and 0.9 respectively). However, we see

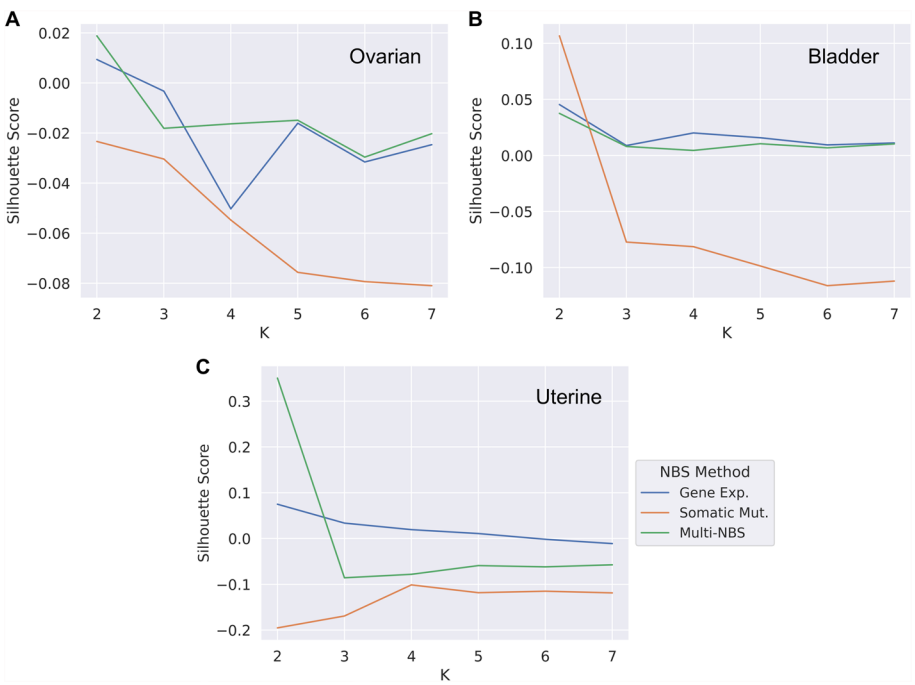

Fig. 2 Silhouette score analysis. Silhouette scores of single-data-type generated clusters and Multi-NBS generated clusters. Refer to the legend in subfigure **C** for all subfigures. **A** Silhouette scores of ovarian cancer clusters. **B** Silhouette scores of bladder cancer clusters. **C** Silhouette scores of uterine cancer clusters

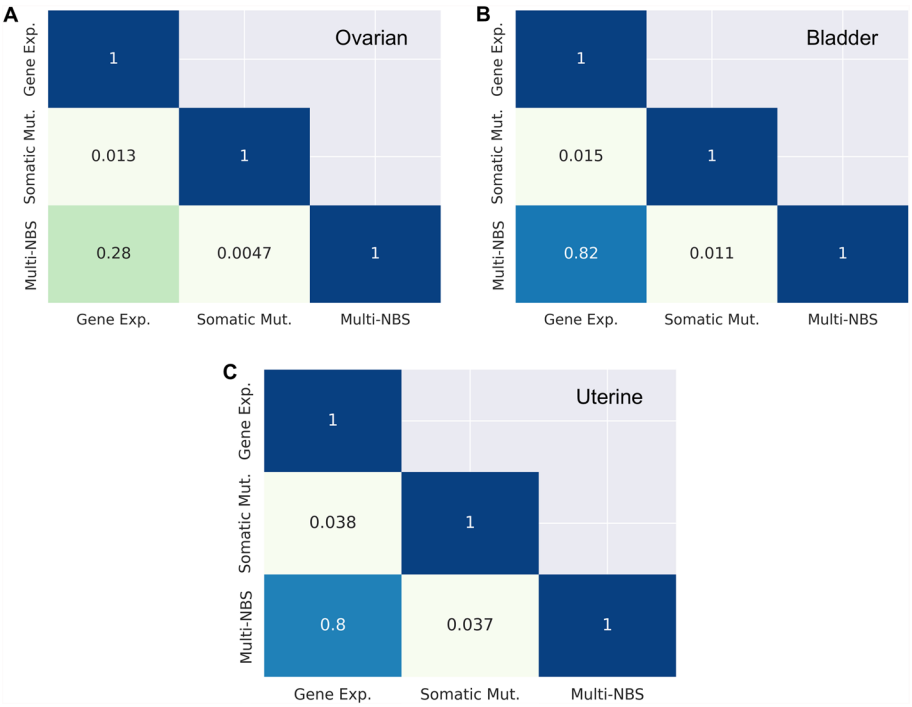

Fig. 3 Adjusted mutual information scores. Adjusted Mutual Information scores of single-data-type generated clusters and Multi-NBS generated clusters. **A** AMI scores for ovarian cancer clusters. **B** AMI scores for bladder cancer clusters. **C** AMI scores for uterine cancer clusters

the AMI scores between Multi-NBS and somatic mutation clusters are disproportionately lower in all three cases, possibly because the somatic mutation profile is binary while RNA-seq profile is continuous such that the integrated profile more resembles the RNA-seq profile. In the following we shall investigate how the blended gene profile may benefit characterization of the cancer subtypes clinically.

Survival analysis

To further assess the generated subtypes, we first examined the Kaplan Meier estimate of the survival function of the generated cancer subtypes, and then a Cox's proportional hazard regression model to investigate how age, gender, and race as covariates may affect survival in addition to the cluster assignments. We found that uterine cancer subtypes, whether identified through single-data type NBS or integrated NBS, did not show significant associations with survival due to the low mortality rate within the cohort at significance level $\alpha = 0.05$ (to be used throughout the following context). For instance, uterine cancer subtypes derived solely from somatic mutation data for $K = 3$ did not exhibit a significant association with survival (Additional file 1: Figure S1, log-rank test, $p = 3.837 \times 10^{-1}$). These findings are consistent with previous work [21]. Consequently, further survival analysis for uterine cancer was not pursued.

In ovarian cancer, we first consider $K = 4$ Multi-NBS subtypes, as it is classified into 4 subtypes based on cancer histology. We note that Multi-NBS subtypes were significantly associated with Kaplan Meier survival functions (Fig. 4A, log-rank test, $p = 6 \times 10^{-3}$ for $K = 4$). In contrast, subtypes derived from somatic mutation or RNA-seq profiles

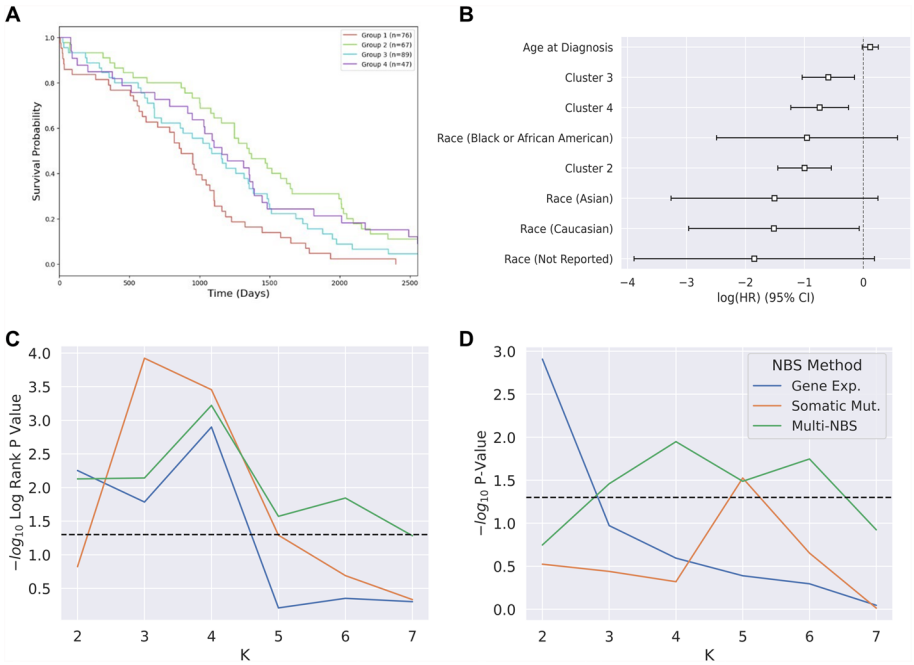

Fig. 4 Ovarian cancer survival analysis. Ovarian cancer survival analysis indicate that integrated subtypes are significantly informative of survival. Refer to the legend in subfigure **D** for subfigures **C** and **D**. **A** Survival curves indicating the probability of survival for subtypes generated by Multi-NBS (log-rank test, $p = 6 \times 10^{-3}$). **B** The Cox regression log hazard ratios for $K = 4$ integrated subtypes. **C** Association of NBS ovarian cancer subtypes to survival time through the log-rank test. The black line represents $p = 0.05$. **D** The predictive power of generated OV subtypes on survival through the log-likelihood ratio test. The black line represents $p = 0.05$

for $K = 4$ were not found to be significantly associated with survival (Additional file 2: Figure S2, log-rank test, $p = 1.79 \times 10^{-1}$ and $p = 1.423 \times 10^{-1}$, respectively). However, we did observe ovarian cancer subtypes generated from somatic mutation data alone for $K = 2$ are significantly associated with survival, consistent with prior work (Additional file 1: Figure S1, log-rank test, $p = 9.8 \times 10^{-3}$) [21, 35]. The difference in the p-value is mostly likely due to the different samples used in each of these studies.

In the Cox-regression model, we found the log hazard ratio for cluster subtypes and Caucasian covariates were all significantly different from zero (Fig. 4B). The observed negative log hazard ratio of the three cluster assignments indicated that subtypes 2 through 4 had a significantly lower hazard compared to subtype 1. This result confirmed the Kaplan Meier estimates of the survival function, where group 1 showed the shortest survival time. In contrast for the single-data type clusters, we found the log hazard ratio for the subtypes did not significantly differ from zero (Additional file 2: Figure S2).

When we varied K for the subtypes definition, Multi-NBS subtypes exhibited more informative survival patterns across various values of K, demonstrating comparable or superior performance in associating subtypes with survival outcomes compared to subtypes derived from single-data types (Fig. 4C). Notably, subtypes derived solely from somatic mutation profiles also show significant associations with Kaplan Meier survival functions, consistent with prior work on somatic mutation NBS subtypes [21]. When we accounted for clinical covariates such as age and race, we found Multi-NBS continued

to consistently yield robust subtypes that remained more predictive of survival than somatic mutation or RNA-seq profiles for $K > 2$ (Fig. 4D). Test statistics, p-values, confidence intervals, and additional statistics for ovarian subtyping survival analysis are specified in (Additional file 3: Table S1).

For bladder cancer, we found that Multi-NBS subtypes proved highly predictive of patient survival time (Fig. 5A, log-rank test, $p = 2 \times 10^{-3}$ for $K = 4$). Notably, the most aggressive bladder cancer Multi-NBS subtype exhibited an average survival time of 1100 days (36 months), while the least aggressive bladder cancer subtype showed an average survival time exceeding 2500 days (82 months). We found subtypes generated using only somatic mutation profiles or RNA-seq profiles were not significantly associated with survival (Additional file 4: Figure S3 log-rank test, $p = 1.989 \times 10^{-1}$ and $p = 7.51 \times 10^{-2}$ for $K = 4$, respectively). Log hazard ratios for $K = 4$ show that all cluster assignments coefficients except for the subtype 2 coefficient significantly differ from zero (Fig. 5B). Similar to ovarian cancer, we see all single-data type cluster assignment coefficients do not significantly differ from zero, indicating that the subtypes generated from mutation or RNA-seq data alone are less informative of survival when covariates are taken into account (Additional file 4: Figure S3). We also see Multi-NBS BLCA subtypes had comparable predictive power to RNA-seq generated subtypes for survival time even after adjusting for clinical covariates such as age, gender, and race (Fig. 5C). This highlights the potential utility and clinical relevance of Multi-NBS in delineating patient subtypes with distinct survival outcomes across cancer types, yielding subtypes that may be more informative of survival compared to single-data type subtypes. P-values,

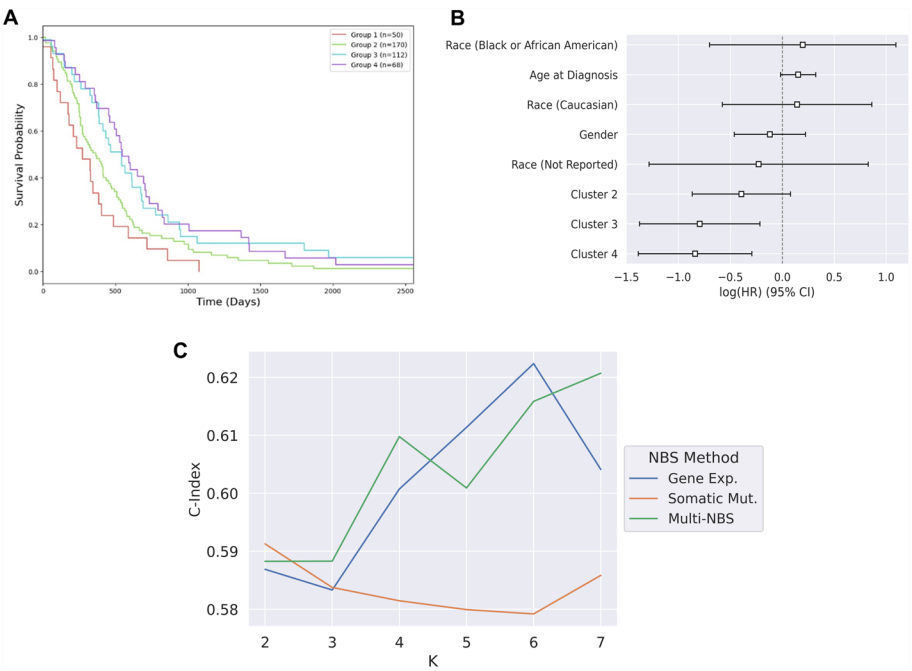

Fig. 5 Bladder cancer survival analysis. Bladder cancer survival analysis indicate that integrated subtypes are significantly informative of survival. **A** Survival curves indicating the probability of survival for subtypes generated by Multi-NBS (log-rank test, $p = 2 \times 10^{-3}$). **B** Cox regression log hazard ratios for $K = 4$ integrated subtypes. **C** Concordance index of different clusters for bladder cancer subtypes

confidence intervals, and additional statistics from our bladder subtyping survival analysis are specified in (Additional file 5: Table S2).

Association with TCGA database subtypes

To further investigate the biological significance, we applied χ^2 association tests with subtypes derived from other data types in the TCGA, including copy-number variation (CNV), methylation, mRNA expression, microRNA expression and protein profiles [44–46]. These TCGA subtypes provided by the TCGA database are determined through analyzing various molecular and histological factors [47–49]. For more information, refer to the Methods section.

For ovarian cancer, we see none of the generated clusters showed a significant association with TCGA subtypes (Fig. 6A). This indicates that the single-data type and Multi-NBS subtypes identified are independent of TCGA subtypes. Cluster similarity analysis between Multi-NBS subtypes and TCGA subtypes (AMI $= -0.0013$) also supports Multi-NBS subtype assignments being independent of TCGA subtypes. For bladder cancer, however, we found both TCGA subtypes and Multi-NBS subtypes are significantly associated with survival, though Multi-NBS is significantly associated with TCGA subtypes more consistently than single-data types with log-rank test, $p = 1.16 \times 10^{-2}$ and $p = 4 \times 10^{-4}$, respectively (Fig. 6B, Additional file 6: Figure S4). For uterine cancer, we see Multi-NBS subtypes exhibited stronger and more consistent association with TCGA subtypes than single-data type clusters (Fig. 6C). We also note

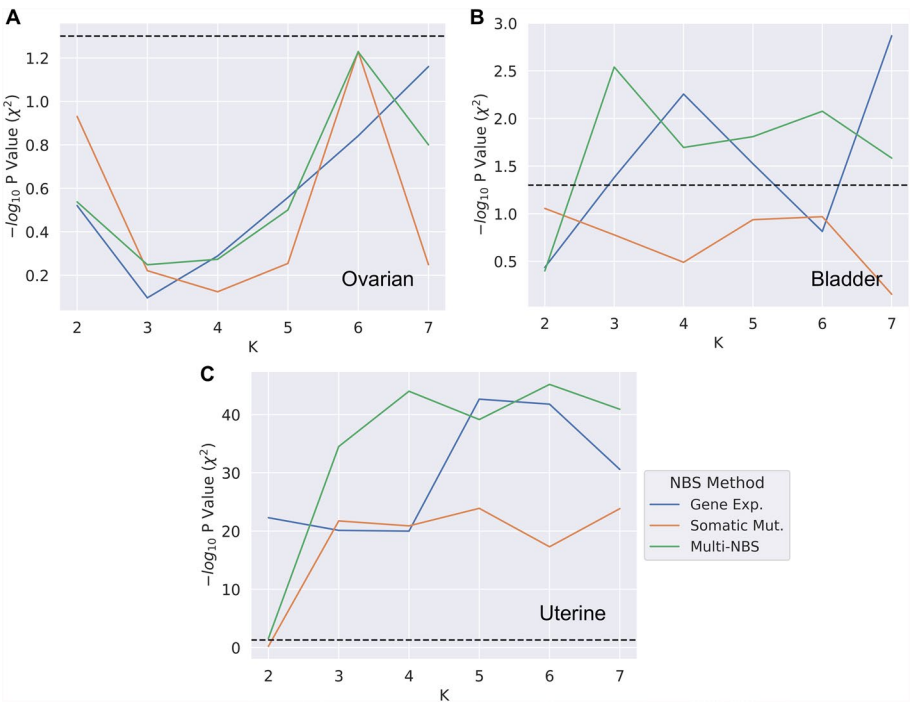

Fig. 6 Association with TCGA subtypes. Association tests with TCGA provided subtypes across various cancers. The dotted line represents $p = 0.05$. Refer to the legend in subfigure **C** for all subfigures. **A** Generated ovarian cancer subtypes association with TCGA ovarian cancer subtypes. **B** Generated bladder cancer subtypes association with TCGA bladder cancer subtypes. **C** Generated uterine cancer subtypes association with TCGA uterine cancer subtypes

that NBS subtypes produced from only somatic mutation profiles are also significantly associated with TCGA subtypes, which is in line with previous findings [21]. Overall, our analysis has revealed that integrated profile-generated subtypes through NBS are potentially more capable than single-data type NBS subtypes of forming stronger association with TCGA subtypes for bladder cancer and uterine cancer.

High-scoring genes

Next, we sought to identify genes with high scores across subtypes generated from Multi-NBS using network smoothed integrated profiles. Since the network values for integrated profile networks are a combination of gene expression and somatic mutation data, we refer to these scores as hybrid scores. We identified the genes with the highest average hybrid score per individual for each subtype. We then compared these findings to high-scoring genes across NBS subtypes generated from only somatic mutation profiles using network smoothed somatic mutation profiles. Again, we identified the genes with the highest average mutation score per individual for each subtype. We limited our investigation to the top 10 highest scoring genes (see Methods for further information on high-scoring genes identification). Across all three cancers, we found that integrated profiles identified more high-scoring genes shared by all subtypes than somatic mutation profiles. We see there were also genes with high scores unique to a subtype or a subset of subtypes.

Here we investigate the genes shared across all subtypes in integrated profiles that were not identified in somatic mutation only profiles to assess their impact on tumor initiation and progression. In ovarian cancer, we found integrated profiles identified BBC3 and UBC as two genes that had high scores across all Multi-NBS subtypes. We see somatic mutation only subtypes only identified NFKBIZ as a high-scoring gene shared by 3 of the 4 somatic mutation NBS subtypes. We found the BBC3 gene had the highest score across all 4 Multi-NBS subtypes. BBC3, or p53 upregulated modulator of apoptosis, whose expression is modulated by tumor suppressor p53. BBC3 has been identified as a strong marker for indicators of cancer [53]. UBC, which has a role in maintaining ubiquitin homeostasis under stress, also has a high score across all subtypes. Current research indicates that UBC is upregulated in cancers and plays a role in the increased proliferation rate of cancer cells and their ability to overcome cellular stresses introduced by cancer treatments [54]. We list the high-scoring genes for integrated NBS subtypes and somatic mutation NBS subtypes for ovarian cancer (Table 1 and Additional file 7: Table S3, respectively).

In bladder cancer, we observe integrated profiles identified ABCC1, LEPR, LTB4R, and IL1RAP as high-scoring genes across all 4 Multi-NBS subtypes while somatic mutation profiles identified ABCC1 and IL1RAP as high-scoring genes present across all 4 NBS subtypes generated from somatic mutation profiles only (Table 2 and Additional file 8: Table S4). LEPR is understood to be involved in tumor proliferation and migration through regulating ERK1/ERK2 and JAK2/STAT3 expression by interacting with ANXA7 [55]. Recently, there has been literature surrounding the possibility of a variant of LEPR being involved in bladder cancer due to the role of leptin in obese individuals [56]. LTB4R, or leukotriene B4 receptors, are involved in cell proliferation, survival, and

Table 1 Integrated ovarian subtypes hybrid scores

Rank	Gene	Subtype	Gene	Subtype	Gene	Subtype	Gene	Subtype
1	BBC3	0.159	BBC3	0.146	BBC3	0.166	BBC3	0.157
2	UBC	0.127	MAPK11	0.072	UBC	0.11	UBC	0.121
3	CCNB1	0.099	UBC	0.071	JAK3	0.097	RB1	0.097
4	YWHAB	0.099	ORC5	0.068	HRAS	0.096	KRAS	0.091
5	DHX9	0.098	NOTCH3	0.068	GSK3A	0.093	HRAS	0.091
6	HSP90AB1	0.097	EIF3E	0.066	IL6	0.091	LIF	0.091
7	TNF	0.096	PPP3R1	0.066	RB1	0.091	JAK3	0.091
8	RB1	0.095	NAP1L1	0.065	SMC3	0.09	MYL7	0.089
9	TERT	0.095	IFNG	0.064	STAT3	0.09	A2M	0.088
10	HRAS	0.092	PPP2R5B	0.064	KRAS	0.09	FN1	0.088

The genes with the highest mean hybrid score across the networks of the patients in ovarian cancer subtypes using integrated profiles. BBC3 and UBC are relevant high-scoring genes across all 4 subtypes

Table 2 Integrated bladder subtypes hybrid scores

Rank	Gene	Subtype	Gene	Subtype	Gene	Subtype	Gene	Subtype
1	ABCC1	0.155	ABCC1	0.157	LEPR	0.09	LTB4R	0.087
2	CCNH	0.105	IL1RAP	0.095	ABCC1	0.077	ABCC1	0.079
3	IL1RAP	0.069	LEPR	0.076	LTB4R	0.067	IL1RAP	0.06
4	AFAP1L2	0.055	CCNH	0.075	IL1RAP	0.06	AFAP1L2	0.049
5	IBSP	0.054	LTB4R	0.066	KLK3	0.057	KLK3	0.044
6	PICALM	0.048	AFAP1L2	0.059	KIR2DS5	0.054	LEPR	0.043
7	THBS3	0.045	PICALM	0.05	ACAD8	0.05	NFE2L2	0.041
8	TICAM2	0.045	HLA-E	0.047	HLA-E	0.046	RBM15	0.038
9	LTB4R	0.045	KIR2DS5	0.046	SF3B1	0.045	PICALM	0.038
10	LEPR	0.042	CXCL1	0.046	NTHL1	0.042	ACAD8	0.036

The genes with the highest mean hybrid score across the networks of the patients in bladder cancer subtypes using integrated profiles. ABCC1, LEPR, LTB4R, and IL1RAP are relevant high-scoring genes across all 4 subtypes

metastasis. Previous literature has identified LTB4R as a potential biomarker for tumor sensitivity [57].

In uterine cancer, we found integrated profiles identified IDO1, NCK2, EDAR, and DUSP8 as high-scoring genes across all 3 Multi-NBS subtypes. We observe somatic mutation profiles only identified IDO1 as a high-scoring gene shared by all 3 NBS subtypes generated from somatic mutation profiles. NCK2 is highly involved in pathways mediating cytoskeleton organization and cell proliferation with previous literature indicating NCK2 expression is linked to metastatic human melanoma tumors [58]. EDAR is part for the Tumor Necrosis Factor Receptor (TNFR) superfamily and is a death receptor. Research has identified EDAR as an oncogene for breast cancer, and could be a potential target for investigation in other cancers [59]. DUSP8 is involved in the mitogen-activated protein kinase (MAPK) signaling pathway and has been shown to be tumor progression and drug resistance in various cancers such as colorectal cancer, lung cancer, and breast cancer [60–62]. These previous findings indicate that DUSP8 could be a potential candidate for its involvement in uterine cancer. We list the high-scoring genes for integrated NBS subtypes and somatic mutation NBS subtypes for uterine cancer (Table 3 and Additional file 9: Table S5, respectively).

Table 3 Integrated uterine subtypes hybrid scores

Rank	Gene	Subtype	Gene	Subtype	Gene	Subtype
1	IDO1	0.148	NCK2	0.203	NCK2	0.165
2	NCK2	0.146	IDO1	0.159	IDO1	0.159
3	ANAPC7	0.125	EDAR	0.129	EDAR	0.136
4	EDAR	0.123	DUSP8	0.117	CDC42	0.111
5	DUSP8	0.083	VWF	0.108	DUSP8	0.107
6	VWF	0.081	MYL12B	0.099	ANAPC7	0.107
7	EGF	0.076	CXCL1	0.084	MYL12B	0.096
8	CD48	0.075	CCNH	0.079	UBC	0.095
9	UBC	0.075	CDC42	0.075	CXCL1	0.095
10	ARID1B	0.074	EGF	0.072	MSH2	0.093

The genes with the highest mean hybrid score across the networks of the patients in uterine cancer subtypes using integrated profiles. IDO1, NCK2, EDAR, and DUSP8 are relevant high-scoring genes across all 3 subtypes

In summary, we found integrated NBS subtypes exhibit a potentially enhanced capability in identifying highly significant genes shared across all subtypes compared to somatic mutation NBS subtypes. The observed shared high-scoring genes are unique for each type of cancer, as evidenced by our analysis which revealed no overlap in these genes among the three cancer types studied. Moreover, the genes within the integrated networks have a documented role in promoting tumor initiation, progression, and drug resistance, as supported by our investigation of existing literature. This underscores the potential of integrated NBS to pinpoint cancer-specific genes warranting further investigation.

Pathway enrichment analysis

We extended our biological analysis by performing pathway enrichment analysis using GSEA, with the KEGG pathway database. The goal was to identify pathways enriched across different subtypes generated from Multi-NBS using the network-smoothed integrated profiles. As a case study, we examined ovarian cancer, while also providing similar analysis results for bladder and uterine cancer.

In ovarian cancer, subtype 4 exhibited the highest number of pathway enrichments, followed by subtype 3 and subtype 2, while subtype 1 showed minimal enrichments (Additional file 10: Table S6). Subtype 2 shows a strong association with cytokine and chemokine receptor interactions, which play a crucial role in immune modulation. These signaling pathways are well-established targets for cancer therapy due to their involvement in tumor development, immune evasion, and proliferation [63, 64]. This suggests that subtype 2 may be primarily driven by up-regulated cytokine and chemokine activity. Subtype 3 exhibits strong associations with the Wnt (β-catenin) and Hippo signaling pathways. Hyperactivation of the Wnt pathway has been observed in ovarian cancer, contributing to cancer stem cell maintenance, metastasis, and chemoresistance [65]. Similarly, amplifications in the Hippo pathway, which regulates tissue homeostasis and development, have been linked to increased metastasis and drug resistance [66]. Subtype 4 demonstrated the most pathway enrichments among the subtypes, with significant hits in the cell cycle and ubiquitin-mediated proteolysis pathways. Dysregulation of the cell cycle is a hallmark of cancer, leading to uncontrolled proliferation [67].

Ubiquitin-mediated proteolysis, responsible for tagging proteins for degradation, regulates key proteins controlling the cell cycle and apoptosis. Disruptions in this pathway can promote tumor growth by preventing apoptosis and sustaining cancer cell survival [68]. Interestingly, subtype 1 only shows one pathway enrichment to protein digestion and absorption.

These distinct patterns of pathway regulation could offer insights into the biological characteristics and therapeutic vulnerabilities of each subtype. Subtype 2 and subtype 3 may represent more aggressive or treatment-resistant forms of cancer, potentially driven by up-regulated signaling activity in immune and developmental pathways. The up-regulated cytokine and chemokine signaling in subtype 2 suggests an inflammatory tumor microenvironment, which could contribute to immune evasion. In contrast, the up-regulated Wnt and Hippo signaling in subtype 3 highlights the role of developmental pathway dysregulation in promoting tumor growth and metastasis. Conversely, the widespread down-regulation of pathways in subtype 4 may suggest a unique tumor biology characterized by disrupted cellular functions and impaired regulatory mechanisms. This could indicate a loss of key signaling pathways necessary for normal cellular control. A similar analysis for bladder cancer and uterine cancer highlighted other pathways relevant in cancer, such as PI3K-Akt signaling and JAK-STAT signaling (Additional file 11: Table S7, Additional file 12: Table S8). Our analysis of enriched pathways and high-scoring genes demonstrates that Multi-NBS is capable of generating clinically significant subtypes, while potentially uncovering biological differences.

Discussion

In this study, we proposed a method that integrates RNA-seq and somatic mutation profiles before undergoing the NBS process to stratify tumors in an unsupervised manner. Through subsequent analysis, we identified the integrated clusters form comparably separated clusters compared to single-data type clusters. We found this integrated approach is also more effective compared to single-data type NBS in producing subtypes associated with survival, even after accounting for clinical covariates. We see this integrated approach also yielded subtypes more significantly associated with TCGA provided subtypes. We also found this integrated approach identified key genes with high mutation scores that were specific to a subtype but also other genes that were highly prevalent across all subtypes. Lastly, we examined enriched pathways that demonstrated potential biological differences between Multi-NBS subtypes.

The reason Multi-NBS performs well could be largely due to the more holistic nature of integrating somatic mutation and RNA-seq profiles. Somatic mutation profiles are a relative comparison of healthy and tumor tissue, while RNA-seq profiles are an absolute measure of the cell state. Somatic mutations are useful because they contain casual genetic signals driving tumor progression and are differential measures between cancer tumors and normal tissue, which is more capable of capturing tumor-specific alterations. However, the tumor phenotype is often variable even when causal genes are identified, and RNA-seq profiles potentially offer to bridge this gap between phenotype and causal genes by better capturing highly expressed genes. Notably, integrated profiles identified more genes with high network scores across all subtypes within a cancer type. Thus, integrated profiles provide a more comprehensive view of the molecular processes

underlying tumor initiation and progression, resulting in the improved stratification we observed. Single-data types such as gene expression are also prone to challenges such as data collection, which can lead to biased results and interpretation [12, 13]. However, these biases can potentially be mitigated through integrative approaches that incorporate multiple layers of genetic data, providing a more holistic view of disease biology to improve predictive accuracy [14, 15, 19]. Rather than evaluating the molecular levels in isolation, integrative network approaches like Multi-NBS provide a more comprehensive framework for studying human health and disease [18, 69]. By more effectively capturing disease impact on biological processes across multiple molecular layers, integrative approaches improve our understanding of cancer biology and provide more informed predictions of disease outcomes and treatment responses.

It is crucial to acknowledge that the selection of β values may vary across cancer types, revealing a potential vulnerability of the method. This variability underscores the importance of tailoring the integration process to the unique characteristics of each cancer type for subtype analysis. In our study, integrating somatic mutation and gene expression profiles involved tuning the hyperparameter β to balance these two data types which led to Multi-NBS subtypes exhibiting improved biological and clinical relevance. However, determining optimal β values proved challenging and varied across cancers. For instance, in uterine cancer, survival analysis did not yield meaningful subtypes, whereas TCGA subtype association metrics did, demonstrating the difficulty of relying on a universal metric to determine the optimal value of β. Another challenge is metric trade-off; an optimal value β for survival analysis may not be ideal for association metrics. For example, selection of β for bladder cancer required us to make ad hoc evaluations. In survival analysis and TCGA subtype association, survival analysis only provided a small subset of β values that were significantly associated with survival (Additional file 13: Figure S5). However, TCGA subtype association suggested a broader range of β values with significant associations. To refine our selection, we focused on evaluating TCGA subtype associations within the subset of β values that were significant in survival analysis. After examining TCGA subtype association within this subset of β values, we concluded that a β value of 0.3 provided the best balance in performance across TCGA subtype association and survival analysis. For ovarian, bladder and uterine cancers, we used tuned β values of 0.8, 0.3 and 0.1, respectively, guided by survival analysis metrics or TCGA subtype association metrics. The fluctuation of β across cancers emphasizes the need for careful consideration and adaptation when implementing Multi-NBS for cancer subtyping to suit the specific biological nuances of each cancer type. Our selection of beta in this study is empirical, as there is currently no single metric to determine the optimal beta value between heterogeneous types of cancer. Future work on developing robust hyperparameter tuning methods for Multi-NBS could reduce the dependency on ad hoc β evaluations.

Various methods exist for integrating multiple genetic data layers within the NBS framework. Multiple additional layers of genetic data such as copy-number variation and protein expression data could be integrated similarly to identify clinically and biologically relevant subtypes. Exploring alternative data integration methods beyond linear combination may enhance stratification. This could include using protein-protein interaction, metabolic, or signaling networks in combination with genetic profiles. Recent

advancements in cancer diagnostics, such as liquid biopsies, methylation signatures, and sequencing systems, have also expanded the range and quality of data available for NBS methods [70–72]. These techniques offer greater sensitivity and robustness against data collection biases, improving the clinical relevance and interpretation of disease subtypes. Additionally, the expanded range of data types presents opportunities for more comprehensive subtyping approaches. By effectively integrating these diverse data types, subtyping methods can achieve higher resolution classifications, enable earlier cancer detection, and maintain consistent performance across various cancer types. Exploration of NBS-generated subtypes through different genetic data integration approaches with these recent advancements could lead to more clinically and biologically informative subtypes, enhancing our understanding of tumor initiation and progression.

Conclusion

In this study, we present a network-based stratification method integrating somatic mutations and gene expression for cancer subtyping. To our knowledge, this specific method of data integration within the NBS framework has not been attempted before, and we concluded that Multi-NBS enhances subtype resolution compared to single-data type analyses. We believe our study offers new insight into data integration in cancer informatics for the purpose of developing personalized medicine and determining biological drivers of cancer.

Abbreviations

TCGA	The Cancer Genome Atlas
ICGC	International Cancer Genome Consortium
NBS	Network-based stratification
NMF	Non-negative matrix factorization
GSEA	Gene set enrichment analysis
FDR	False discovery rate
KEGG	Kyoto Encyclopedia of Genes and Genomes
RNA	Ribonucleic acid
AMI	Adjusted mutual information

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06143-y.

Additional file 1: Fig. S1. Somatic mutation generated subtypes. NBS subtypes using only somatic mutation profiles for ovarian cancer ($K = 2$) and uterine cancer ($K = 3$) with the associated survival log-rank test p-value.

Additional file 2: Fig. S2. Ovarian cancer single-data type survival analysis. (A, C) Ovarian cancer single-data type survival curves for $K = 4$ with the log-rank test p-value above each survival curve. Single-data type subtypes are not significantly associated with survival while integrated profile generated subtypes are significantly associated with survival as shown in the paper. (B, D) Ovarian cancer log hazard ratios for single-data type generated clusters. The cluster coefficients are not significantly nonzero indicating the cluster assignment is not informative of higher or lower survival than subtype 1.

Additional file 3: Table S1. Ovarian subtype survival analysis statistics. Test statistics, p-values, confidence intervals, and additional statistics are provided.

Additional file 4: Fig. S3. Bladder cancer single-data type survival analysis. (A, C) Bladder cancer single-data type survival curves for $K = 3$ with the log-rank test p-value above each survival curve. Single-data type subtypes are not significantly associated with survival while integrated profile generated subtypes are significantly associated with survival as shown in the paper. (B, D) Bladder cancer log hazard ratios for single-data type generated clusters. The cluster coefficients are not significantly nonzero indicating the cluster assignment is not informative of higher or lower survival than subtype 1.

Additional file 5: Table S2. Bladder subtype survival analysis statistics. Confidence intervals, p-values, and additional statistics are provided.

Additional file 6: Fig. S4. TCGA subtypes and Multi-NBS subtype survival curves. Survival curves from ovarian cancer TCGA subtypes and Multi-NBS generated subtypes ($K = 4$). Both are significantly associated with survival through log-rank test.

Additional file 7: Table S3. Somatic mutation ovarian subtypes mutation scores. The genes with the highest mean mutation score across the networks of the patients in ovarian cancer subtypes using only somatic mutation profiles profiles. NFKBIZ is a relevant high-scoring genes across 3 of the 4 subtypes.

Additional file 8: Table S4. Somatic mutation bladder subtypes mutation scores. The genes with the highest mean mutation score across the networks of the patients in bladder cancer subtypes using only somatic mutation profiles profiles.. ABCC1 and IL1RAP are relevant high-scoring genes across all 4 subtypes. LEPR is a high scoring gene across 3 of the 4 subtypes.

Additional file 9: Table S5. Somatic mutation uterine subtypes mutation scores. The genes with the highest mean mutation score across the networks of the patients in uterine cancer subtypes using only somatic mutation profiles. IDO1 is a relevant high-scoring gene across all 3 subtypes. NCK2, EDAR, and DUSP8 are high-scoring genes in 2 out of the 3 subtypes.

Additional file 10: Table S6. Enriched pathways for Multi-NBS ovarian cancer subtypes. Enriched pathways for each subtype are labeled and include output statistics and annotations.

Additional file 11: Table S7. Enriched pathways for Multi-NBS bladder cancer subtypes. Enriched pathways for each subtype are labeled and include output statistics and annotations.

Additional file 12: Table S8. Enriched pathways for Multi-NBS uterine cancer subtypes. Enriched pathways for each subtype are labeled and include output statistics and annotations.

Additional file 13: Fig. S5. β hyperparameter tuning for bladder cancer. (A) TCGA subtype association $-log_{10}$ p-values for various values of β across K number of subtypes. (B) KM survival analysis $-log_{10}$ p-values for various values of β across K number of subtypes.

Acknowledgements
The authors thank the anonymous referees for their thoughtful input and contributions for strengthening this manuscript.

Author contributions
JW conceived the study. JW and KL designed the study. KL implemented the algorithm and performed the analysis. JW and KL interpreted the results. JW and KL wrote the manuscript. All authors read and approved the final manuscript for publication.

Funding
The research reported in this paper was supported by the NSF Quantitative Biology REU Site at Northwestern (DMS-2150134); Northwestern Research Training Grant in Quantitative Biological Modeling, (NSF DMS-1547394); NSF-Simons Center for Quantitative Biology (Simons Foundation/SFARI 597491-RWC and NSF DMS-1764421). The authors gratefully acknowledge the support provided by these grants.

Data availability
The example datasets and notebooks along with an implementation of Multi-NBS from the current study are available in the Multi-NBS repository.

Declarations

Ethics approval and consent to participate
We confirm that all methods and procedures conducted in this study were carried out in accordance with the relevant guidelines and regulations set forth by The Cancer Genome Atlas in regards to human subjects protection and data access policies.

Consent for publication
Not applicable.

Competing Interests
The authors declare no Competing Interest.

Received: 28 November 2024 Accepted: 15 April 2025
Published online: 13 May 2025

References

1. The International Cancer Genome Consortium. International network of cancer genome projects. Nature. 2010;464(7291):993–8. https://doi.org/10.1038/nature08987.

2. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609–15. https://doi.org/10.1038/nature10166.
3. Liu H, Tang T. Pan-cancer genetic analysis of disulfidptosis-related gene set. Cancer Genet. 2023;278–279:91–103. https://doi.org/10.1016/j.cancergen.2023.10.001.
4. Rasteh AM, Liu H, Wang P. Pan-cancer genetic profiles of mitotic DNA integrity checkpoint protein kinases. Cancer Biomark. 2025. https://doi.org/10.3233/CBM-240119.
5. Liu H, Weng J, Huang CL, Jackson AP. Is the voltage-gated sodium channel β3 subunit (SCN3B) a biomarker for glioma? Funct Integr Genom. 2024;24(5):162. https://doi.org/10.1007/s10142-024-01443-7.
6. Liu H, Tang T. MAPK signaling pathway-based glioma subtypes, machine-learning risk model, and key hub proteins identification. Sci Rep. 2023;13:19055. https://doi.org/10.1038/s41598-023-45774-0.
7. Whiteside TL. The tumor microenvironment and its role in promoting tumor growth. Oncogene. 2008;27(45):5904–12. https://doi.org/10.1038/onc.2008.271.
8. Konstantinopoulos PA, Spentzos D, Cannistra SA. Gene expression profiling in epithelial ovarian cancer. Nat Rev Clin Oncol. 2008;5(10):577–87. https://doi.org/10.1038/ncponc1178.
9. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet. 2011;378(9805):1812–23. https://doi.org/10.1016/S0140-6736(11)61539-0.
10. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330–7. https://doi.org/10.1038/nature11252.
11. Raspe E. Gene expression profiling to dissect the complexity of cancer biology: pitfalls and promise. Semin Cancer Biol. 2012;22(3):250–60. https://doi.org/10.1016/j.semcancer.2012.02.011.
12. Liu H, Guo Z, Wang P. Genetic expression in cancer research: challenges and complexity. Gene Rep. 2024;37: 102042. https://doi.org/10.1016/j.genrep.2024.102042.
13. Liu H, Li Y, Karsidag M, Tu T, Wang P. Technical and biological biases in bulk transcriptomic data mining for cancer research. J Cancer. 2025;16(1):34–43. https://doi.org/10.7150/jca.100922.
14. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood C, Beyene J. Data integration in genetics and genomics: methods and challenges. Hum Genom Proteom. 2009;2009: 869093. https://doi.org/10.4061/2009/869093.
15. Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. Front Oncol. 2020;10:1030. https://doi.org/10.3389/fonc.2020.01030.
16. Yang G, Xuequn S, Zhanhuai L. Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. Neurocomputing. 2019;324:20–30. https://doi.org/10.1016/j.neucom.2018.03.072.
17. Qianxing M, Xuequn S, Zhanhuai L. Pattern discovery and cancer gene identification in integrated cancer genomic data. PNAS. 2013;110(11):4245–50. https://doi.org/10.1073/pnas.1208949110.
18. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051. https://doi.org/10.1177/1177932219899051.
19. Chen C, Wang J, Pan D, Wang X, Xu Y, Yan J, et al. Applications of multi-omics analysis in human diseases. Med Commun. 2023;4(4): e315. https://doi.org/10.1002/mco2.315.
20. Kreeger PK, Lauffenburger DA. Cancer systems biology: a network modeling perspective. Carcinogenesis. 2010;31(1):2–8. https://doi.org/10.1093/carcin/bgp261.
21. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013;10(11):1108–15. https://doi.org/10.1038/nmeth.2651.
22. Kang L, Liu A, Tian L. Linear combination methods to improve diagnostic/prognostic accuracy on future observations. Stat Methods Med Res. 2013;25(4):1359–80. https://doi.org/10.1177/0962280213481053.
23. Etzioni R, Kooperberg C, Pepe M, Smith R, Gann PH. Combining biomarkers to detect disease with application to prostate cancer. Biostatistics. 2003;4(4):523–38. https://doi.org/10.1093/biostatistics/4.4.523.
24. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. J Am Stat Assoc. 1993;88(424):1350–5. https://doi.org/10.1080/01621459.1993.10476417.
25. Huang JK, Carlin DE, Yu MK, Zhang W, Kresiburg JF, Tamayo P, et al. Systematic evaluation of molecular networks for discovery of disease genes. Cell Syst. 2018;6(4):484-95.e5. https://doi.org/10.1016/j.cels.2018.03.001.
26. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015. https://doi.org/10.1093/nar/gku1075.
27. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74. https://doi.org/10.1016/j.cell.2011.02.013.
28. Iorio F, Knijnenburg TA, Vis DJ, Bignell G, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell. 2016;166(3):740–54. https://doi.org/10.1016/j.cell.2016.06.017.
29. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58. https://doi.org/10.1126/science.1235122.
30. Huang JK, Jia T, Carlin DE, Ideker T. pyNBS: a Python implementation for network-based stratification of tumor mutations. Bioinformatics. 2018;34(16):2859–61. https://doi.org/10.1093/bioinformatics/bty186.
31. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol. 2010. https://doi.org/10.1371/journal.pcbi.1000641.
32. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91. https://doi.org/10.1038/44565.
33. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad. 2004;101(12):4164–9. https://doi.org/10.1073/pnas.0308531101.
34. Cai et al. Non-negative matrix factorization on manifold. In: 8th IEEE international conference in data mining. 2008; Pisa, Italy. p. 63–72. https://doi.org/10.1109/ICDM.2008.57.
35. Zhong X, Yang H, Zhao S, Shyr Y, Li B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. BMC Genom. 2015. https://doi.org/10.1186/1471-2164-16-S7-S7.
36. He Z, Zhang J, Yuan X, Liu Z, Liu B, Tuo S, et al. Network based stratification of major cancers by integrating somatic mutation and gene expression data. PLoS One. 2017;12(5): e0177662. https://doi.org/10.1371/journal.pone.0177662.

37. Monti S, et al. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52:91–118. https://doi.org/10.1023/A:1023949509487.
38. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Appl Comput Math. 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
39. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform. 2014;8:14. https://doi.org/10.3389/fninf.2014.00014.
40. Davidson-Pilon C. Lifelines: survival analysis in Python. J Open Source Softw. 2019;4(40):1317. https://doi.org/10.21105/joss.01317.
41. Pölsterl S, Gupta P, Wang L, Conjeti S, Katouzian A, Navab N. Heterogeneous ensembles for predicting survival of metastatic castrate-resistant prostate cancer patients. F1000Research. 2016;5:2676. https://doi.org/10.12688/f1000research.8231.1.
42. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81. https://doi.org/10.1080/01621459.1958.10501452.
43. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Ann Stat. 1982;10(4):1100–20. https://doi.org/10.1214/aos/1176345976.
44. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016. https://doi.org/10.1093/nar/gkv1507.
45. Silva TC, Colaprico A, Olsen C, D'Angelo F, Bontempi G, Ceccarelli M, et al. TCGA Workflow: analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Research. 2016;5:1542. https://doi.org/10.12688/f1000research.8923.2.
46. Mounir M, Lucchetta M, Silva TC, Olsen C, Bontempi G, Chen X, et al. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS Comput Biol. 2019. https://doi.org/10.1371/journal.pcbi.1006701.
47. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7351):609–15. https://doi.org/10.1038/nature10166.
48. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507(7493):315–22. https://doi.org/10.1038/nature12965.
49. Levine DA. The cancer genome atlas research network. integrated genomic characterization of endometrial carcinoma. Nature. 2013;497(7447):67–73. https://doi.org/10.1038/nature12113.
50. Yu G. Thirteen years of clusterProfiler. Innovation. 2024;5(6): 100722. https://doi.org/10.1016/j.xinn.2024.100722.
51. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation. 2021;2(3): 100141. https://doi.org/10.1016/j.xinn.2021.100141.
52. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27.
53. Han J, Flemington C, Houghton AB, Gu Z, Zambetti GP, Lutz RJ, et al. Expression of bbc3, a pro-apoptotic BH3-only gene, is regulated by diverse cell death and survival signals. PNAS. 2001;98(20):11318–23. https://doi.org/10.1073/pnas.201208798.
54. Dingding S, Grossman SR. Ubiquitin becomes ubiquitous in cancer. Cancer Biol Ther. 2010;10(8):737–47. https://doi.org/10.4161/cbt.10.8.13417.
55. Huang H, Zhang J, Ling F, Huang Y, Yang M, Zhang Y, et al. Leptin receptor (LEPR) promotes proliferation, migration, and invasion and inhibits apoptosis in hepatocellular carcinoma by regulating ANXA7. Cancer Cell Int. 2021;21(4):4. https://doi.org/10.1186/s12935-020-01641-w.
56. Alfaqih MA, Elsalem L, Nusier M, Mhedat K, Khader Y, Ababneh E. Serum leptin receptor and the rs1137101 variant of the LEPR gene are associated with bladder cancer. Biomolecules. 2023;13(10):1498. https://doi.org/10.3390/biom13101498.
57. Kalinkin AI, Nemtsova MV, Zaletaev DV, Sigin VO, Ignatova E, Kuznetsova EB, et al. Leukotriene B4 receptors abnormal gene expression is associated with either shorter or longer survival in breast cancer patients depending on the intrinsic tumour molecular subtype. Ann Oncol. 2019. https://doi.org/10.1093/annonc/mdz413.061.
58. Labelle-Cote M, Dusseault J, Salma I, Picard-Cloutier A, Siegel PM, Larose L. Nck2 promotes human melanoma cell proliferation, migration and invasion in vitro and primary melanoma-derived tumor growth in vivo. BMC Cancer. 2011;11:443. https://doi.org/10.1186/1471-2407-11-443.
59. Williams R, Jobling S, Sims AH, Mou C, Wilkinson L, Collu GM, et al. Elevated EDAR signalling promotes mammary gland tumourigenesis with squamous metaplasia. Oncogene. 2022;41(10):1040–9. https://doi.org/10.1038/s41388-021-01902-6.
60. Zhang H, Wang M, Chen D, Luo C. Dual-specificity phosphatase 8 (DUSP8) induces drug resistance in breast cancer by regulating MAPK pathways. J Investig Med. 2022;70(5):1293–300. https://doi.org/10.1136/jim-2021-002282.
61. Turkowski K, Herzberg F, Günther S, Weigert A, Haselbauer T, Fink L, et al. miR-147b mediated suppression of DUSP8 promotes lung cancer progression. Oncogene. 2024;43(2):1178–89. https://doi.org/10.1038/s41388-024-02969-7.
62. Ding T, Zhou Y, Long R, Chen C, Zhao J, Cui P, et al. DUSP8 phosphatase: structure, functions, expression regulation and the role in human diseases. Cell Biosci. 2019;9:70. https://doi.org/10.1186/s13578-019-0329-4.
63. Yi M, Li T, Niu M, Zhang H, Wu Y, Wu K, et al. Targeting cytokine and chemokine signaling pathways for cancer therapy. Sig Transduct Target Ther. 2024;9:176. https://doi.org/10.1038/s41392-024-01868-3.
64. Abdul-Rahman T, Ghosh S, Badar SM, Nazir A, Bamigbade GB, Aji N, et al. The paradoxical role of cytokines and chemokines at the tumor microenvironment: a comprehensive review. Eur J Med Res. 2024;29:124. https://doi.org/10.1186/s40001-024-01711-z.
65. Nguyen V, Hough R, Bernaudo S, et al. Wnt/beta-catenin signalling in ovarian cancer: insights into its hyperactivation and function in tumorigenesis. J Ovarian Res. 2019;12:122. https://doi.org/10.1186/s13048-019-0596-z.
66. Balakrishnan K, Chen Y, Dong J. Amplification of hippo signaling pathway genes is governed and implicated in the serous subtype-specific ovarian carcino-genesis. Cancers (Basel). 2024;16(9):1781. https://doi.org/10.3390/cancers16091781.

67. Matthews HK, Bertoli C, de Bruin R. Cell cycle control in cancer. Nat Rev Mol Cell Biol. 2022;23:74–88. https://doi.org/10.1038/s41580-021-00404-3.
68. Rao Z, Ding Y. Ubiquitin pathway and ovarian cancer Curr Oncol. 2012;19(6):324–8. https://doi.org/10.3747/co.19.1175. Erratum. In: Curr Oncol. 2013;20(3):e280.
69. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. Brief Bioinform. 2018;19(6):1370–81. https://doi.org/10.1093/bib/bbx066.
70. Jahangiri L. Updates on liquid biopsies in neuroblastoma for treatment response, relapse and recurrence assessment. Cancer Genet. 2024;288–289:32–9. https://doi.org/10.1016/j.cancergen.2024.09.001.
71. Ohyama H, Hirotsu Y, Amemiya K, Mikata R, Amano H, Hirose S, et al. Development of a molecular barcode detection system for pancreaticobiliary malignancies and comparison with next-generation sequencing. Cancer Genet. 2024;280–281:6–12. https://doi.org/10.1016/j.cancergen.2023.12.002.
72. Gonzalez T, Nie Q, Chaudhary LN, Basel D, Reddi HV, et al. Methylation signatures as biomarkers for non-invasive early detection of breast cancer: a systematic review of the literature. Cancer Genet. 2024;282–283:1–8. https://doi.org/10.1016/j.cancergen.2023.12.003.

Publisher's Note