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

Kenny Liou[1*], Ji-Ping Wang[1**]

1 Department of Statistics and Data Science, Northwestern University, Evanston, Illinois, United States of America

* kliou@ini.usc.edu ** jzwang@northwestern.edu

Abstract

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. These studies provide 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 integrate somatic mutation data with RNA sequencing data within the NBS framework and investigate the effectiveness of integrated NBS on three cancers: ovarian, bladder, and uterine cancer. We show that integrated NBS subtypes are more significantly associated with overall survival or histology. Integrated NBS networks also reveal highly influential genes that drive cancer initiation and progression. This comprehensive approach underscores the significance of integrating genomic data types in cancer subtyping, offering profound implications for personalized prognosis and treatment strategies.

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 like the International Cancer Genome Consortium (ICGC) [1] and The Cancer Genome Atlas (TCGA) [2] meticulously profile genomic, transcriptomic, and epigenomic data. This includes DNA methylation, microRNA expression, and protein expression data, unveiling a treasure trove of information. 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 [3].

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 [4] and breast cancer [5]. However, it has fallen short in other cancer types, such as colorectal cancer, where the association between molecular subtypes and clinical phenotypes remains elusive [6]. This limitation may be attributed to issues like sample quality and overfitting, inherent to gene expression analysis [7].

As the landscape of genome-scale data diversifies and expands, there emerges an urgent need to develop methodologies that effectively integrate multi-omics data for more robust tumor subtyping [8]. Multi-omic research in cancer genetics offers promising results, such as in breast cancer [9, 10]. 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 [8]. The advantage of multi-omics is that the potential interactions between various molecular layers are considered, providing a holistic perspective that can lead to different results compared to single-data-type methods [11]. In this paper, we integrate molecular profiles before processing these data types to explore if pre-process integration can yield meaningful results.

Recent years have witnessed the widespread adoption of gene interaction networks,

with cancer being increasingly recognized as a network-driven disease [12]. Somatic mutation profiles harbor cancer-driver genes capable of instigating genetic mutations in other genes. Network-based Stratification (NBS) [13] 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 line with these trends in cancer genomics, we propose a multi-omics methodology grounded in NBS for tumor stratification, fusing somatic mutation profiles with RNA gene expression data. 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 significant, and clinically informative 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.

Materials and 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 data portal in 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. 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. In the integration process of NBS for ovarian cancer, bladder cancer, and uterine cancer, 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 significant p-value 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 www.cancer.gov/ccg/research/genome-sequencing/tcga.

Gene interaction network

The constructed network is derived from PCNet [16], 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: [17–20]. 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 [15].

Network propagation

The integrated profiles are mapped onto the gene interaction network and then network propagation [44] 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 [13]. 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 [45–47]. 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 [13].

Consensus clustering

We used consensus clustering [48] 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 [13]. 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 [15].

Cluster analysis

We used Silhouette scores [21] 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) [22] to assess cluster similarity between clusters formed through different data-type generated clusters.

Survival analysis

Survival analysis was performed through the lifelines [49]and scikit-survival package [50]. Kaplan-Meier survival curves [51] 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 [52]. 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 [23–25], which are determined through genomic, transcriptomic, and proteomic characterization of tumors using array and sequence-based technologies by The Cancer Genome Atlas Research Network [26–28].

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.

Results

For the three cancer types under consideration, ovarian, uterine, and bladder carcinoma, only individules 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 1). The resulting integrated matrix is then processed through NBS to generate patient subtype assignments. The underlying network used in this context is a sub-network comprised of cancer-related genes, which was constructed from PCNet [16], but filtered using four databases containing cancer-specific genes and interactions [17–20]. 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).

Fig 1. Integrated NBS workflow A workflow of an integrated approach to network-based stratification.

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 first evaluated whether the clusters generated by Multi-NBS are comparable to single-data-type clusters using Silhouette scores [21]. 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 good 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 slightly more well-formed clusters (Fig 2A and Fig 2B), while RNA-seq seemed to generate the most well-formed clusters (Fig 2C). The Multi-NBS clusters did not show a uniformly better pattern than 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.

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.

We next compared the similarity of cluster contents using Adjusted Mutual Information (AMI) [22]. 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. We observed that across all cancer types, somatic mutation and RNA-seq clusters were mostly independent of each other (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, 0.8 and 0.9 respectively). However, we see the AMI scores between Multi-NBS and somatic mutation clusters are unproportionally 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.

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.

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 to 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 (S3 Fig, log-rank test, $p = 3.837 \times 10^{-1}$). These findings are consistent with previous work [13]. 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 for $K = 4$ were not found to be significantly associated with survival (S1 Fig, 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 (S3 Fig, log-rank test, $p = 9.8 \times 10^{-3}$) [13, 14]. The difference in the p-value is mostly likely due to the different samples used in each of these studies.

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 $p = 0.05$.

In the Cox-regression model, we found the log hazard ratio for cluster subtypes and Caucasian covariates were all significantly different from zero in 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 (S1 Fig).

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 [13]. 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).

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 (S2 Fig 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 (S2 Fig). 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 in 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.

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.

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 [23–25]. These TCGA subtypes provided by the TCGA database are determined through analyzing various molecular and histological factors [26–28]. 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 in 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, S4 Fig). 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 that NBS subtypes produced from only somatic mutation profiles are also significantly associated with TCGA subtypes, which is in line with previous findings [13]. 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.

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.

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 tumour 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 [29]. 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 [30]. We list the high-scoring genes for integrated NBS

subtypes and somatic mutation NBS subtypes for ovarian cancer in S1 Table and S2 Table, 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 (S3 Table and S4 Table). LEPR is understood to be involved in tumour proliferation and migration through regulating ERK1/ERK2 and JAK2/STAT3 expression by interacting with ANXA7 [31]. 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 [32]. LTB4R, or leukotriene B4 receptors, are involved in cell proliferation, survival, and metastasis. Previous literature has identified LTB4R as a potential biomarker for tumour sensitivity [33].

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 tumours [34]. EDAR is part for the Tumour 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 [35]. DUSP8 is involved in the mitogen-activated protein kinase (MAPK) signaling pathway and has been shown to be tumour progression and drug resistance in various cancers such as colorectal cancer, lung cancer, and breast cancer [36–38]. 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 in S5 Table and S6 Table, respectively.

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.

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.

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, we see integrated profiles were able to identify more genes that had 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.

It is crucial to acknowledge that the choice of β values may not remain consistent

across all cancers, 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 ovarian, bladder, and uterine cancers, we utilized 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.

There are still many methods to integrate different layers of genetic data 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. Another avenue of exploration is the integration of different genetic data that is beyond a linear combination of genetic profiles. This could include using protein-protein interaction, metabolic, or signaling networks in combination with genetic profiles. Exploration of NBS-generated subtypes through different genetic data integration methods could lead to more clinically and biologically informative subtypes. Investigating these integrated NBS approaches across a wide variety of cancers can lead to a stronger understanding of tumor initiation and progression.

Acknowledgments

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.

References

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

2. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7353):609-15. doi: 10.1038/nature10166.

3. Whiteside TL. The tumor microenvironment and its role in promoting tumor growth. *Oncogene*. 2008;27(45):5904-12. doi: 10.1038/onc.2008.271.

4. Konstantinopoulos PA, Spentzos D, Cannistra SA. Gene expression profiling in epithelial ovarian cancer. *Nat Rev Clin Oncol*. 2008;5(10):577-87. doi: 10.1038/ncponc1178.

5. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet*. 2011;378(9805):1812-23. doi: 10.1016/S0140-6736(11)61539-0.

6. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-7. doi: 10.1038/nature11252.

7. Raspe E. Gene expression profiling to dissect the complexity of cancer biology: pitfalls and promise. *Semin Cancer Biol*. 2012;22(3):250-60. doi: 10.1016/j.semcancer.2012.02.011.

8. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J. Data Integration in Genetics and Genomics: Methods and Challenges. *Hum Genomics Proteomics*. 2009;2009:869093. doi: 10.4061/2009/869093.

9. 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. doi: 10.1016/j.neucom.2018.03.072.

10. Qianxing M Xuequn S, Zhanhuai L. Pattern discovery and cancer gene identification in integrated cancer genomic data. *PNAS*. 2013;110(11):4245-50. doi: 10.1073/pnas.1208949110.

11. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights*. 2020;14:1177932219899051. doi: 10.1177/1177932219899051.

12. Kreeger PK, Lauffenburger DA. Cancer systems biology: a network modeling perspective. *Carcinogenesis*. 2010;31(1):2-8. DOI: 10.1093/carcin/bgp261.

13. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat Methods*. 2013;10(11):1108-15. doi: 10.1038/nmeth.2651.

14. 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 Genomics*. 2015;16(Suppl 7). doi: 10.1186/1471-2164-16-S7-S7.

15. 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. doi: 10.1093/bioinformatics/bty186.

16. 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 Systems*. 2018;6(4):484-95.e5. doi: 10.1016/j.cels.2018.03.001.

17. 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;43(D1). doi: 10.1093/nar/gku1075.

18. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-74. doi: 10.1016/j.cell.2011.02.013.

19. 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. doi: 10.1016/j.cell.2016.06.017.

20. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. Cancer genome landscapes. *Science*. 2013;339(6127):1546-58. doi: 10.1126/science.1235122.

21. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Appl Comput Math*. 1987;20:53-65. doi: 10.1016/0377-0427(87)90125-7.

22. 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. doi: 10.3389/fninf.2014.00014.

23. 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;44(8). doi: 10.1093/nar/gkv1507.

24. 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. doi: 10.12688/f1000research.8923.2.

25. 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;15(3). doi: 10.1371/journal.pcbi.1006701.

26. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7351):609-15. doi: 10.1038/nature10166.

27. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*. 2014;507(7493):315-22. doi: 10.1038/nature12965.

28. Levine DA, The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature*. 2013;497(7447):67-73. DOI: 10.1038/nature12113.

29. 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. DOI: 10.1073/pnas.201208798.

30. Dingding S, Grossman SR. Ubiquitin becomes ubiquitous in cancer. *Cancer Biol Ther*. 2010;10(8):737-47. doi: 10.4161/cbt.10.8.13417.

31. 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. doi: 10.1186/s12935-020-01641-w.

32. 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. doi: 10.3390/biom13101498.

33. 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;30(Suppl 7) doi: 10.1093/annonc/mdz413.061.

34. Labelle-Cote M, Dusseault J, Ismaïl Salma, 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. doi: 10.1186/1471-2407-11-443.

35. 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. doi: 10.1038/s41388-021-01902-6.

36. 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. doi: 10.1136/jim-2021-002282.

37. 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. doi: 10.1038/s41388-024-02969-7.

38. 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. doi: 10.1186/s13578-019-0329-4.

39. Wang X, Yong L. Tumor necrosis factor and cancer, buddies or foes? *Acta Pharmacol Sin*. 2008;29(10):1275-88. doi: 10.1111/j.1745-7254.2008.00889.x.

40. Katopodis P, Kerslake R, Zikopoulos A, Beri N, Anikin V. p38B - MAPK11 and its role in female cancers. *J Ovarian Res*. 2021;14:84. doi: 10.1186/s13048-021-00834-9.

41. Huang L, Guo Z, Wang F, Fu L. KRAS mutation: from undruggable to druggable in cancer. *Signal Transduct Target Ther*. 2021;6:386. doi: 10.1038/s41392-021-00780-4.

42. Wu XY, Liu WT, Wu ZF, Chen C, Liu JY, Wu GN et al. Identification of HRAS as cancer-promoting gene in gastric carcinoma cell aggressiveness. *Am J Cancer Res*. 2016;6(9):1935-48. PMID: 27725900; PMCID: PMC5043104.

43. Pazik M, Michalska K, Żebrowska-Nawrocka M, Zawadzka I, Łochowski M, Balcerczak E. Clinical significance of HRAS and KRAS genes expression in patients with non–small-cell lung cancer - preliminary findings. *BMC Cancer*. 2021;21:130. doi: 10.1186/s12885-021-07858-w.

44. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS Comput Biol*. 2010;6(1) doi: 10.1371/journal.pcbi.1000641.

45. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999;401(6755):788–91. doi: 10.1038/44565.

46. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad*. 2004;101(12):4164–9. doi: 10.1073/pnas.0308531101.

47. Cai et al. Non-negative matrix factorization on manifold. *8th IEEE Int. Conf. Data Mining*. 2008; Pisa, Italy. p. 63–72. doi: 10.1109/ICDM.2008.57.

48. 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. doi: 10.1023/A:1023949509487.

49. Davidson-Pilon C. lifelines: survival analysis in Python. *J Open Source Softw.* 2019;4(40):1317. doi: 10.21105/joss.01317.

50. 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. doi: 10.12688/f1000research.8231.1.

51. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc.* 1958;53(282):457–81. doi: 10.1080/01621459.1958.10501452.

52. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat.* 1982;10(4):1100–20. doi: 10.1214/aos/1176345976.

Supporting information

S1 Fig. 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.

S2 Fig. 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.

S3 Fig. 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.

S4 Fig. 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.

S1 Table. Integrated ovarian subtypes hybrid scores. 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.

S2 Table. 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.

S3 Table. Integrated bladder subtypes hybrid scores. 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.

S4 Table. 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.

S5 Table. Integrated uterine subtypes hybrid scores. 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.

S6 Table. 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.

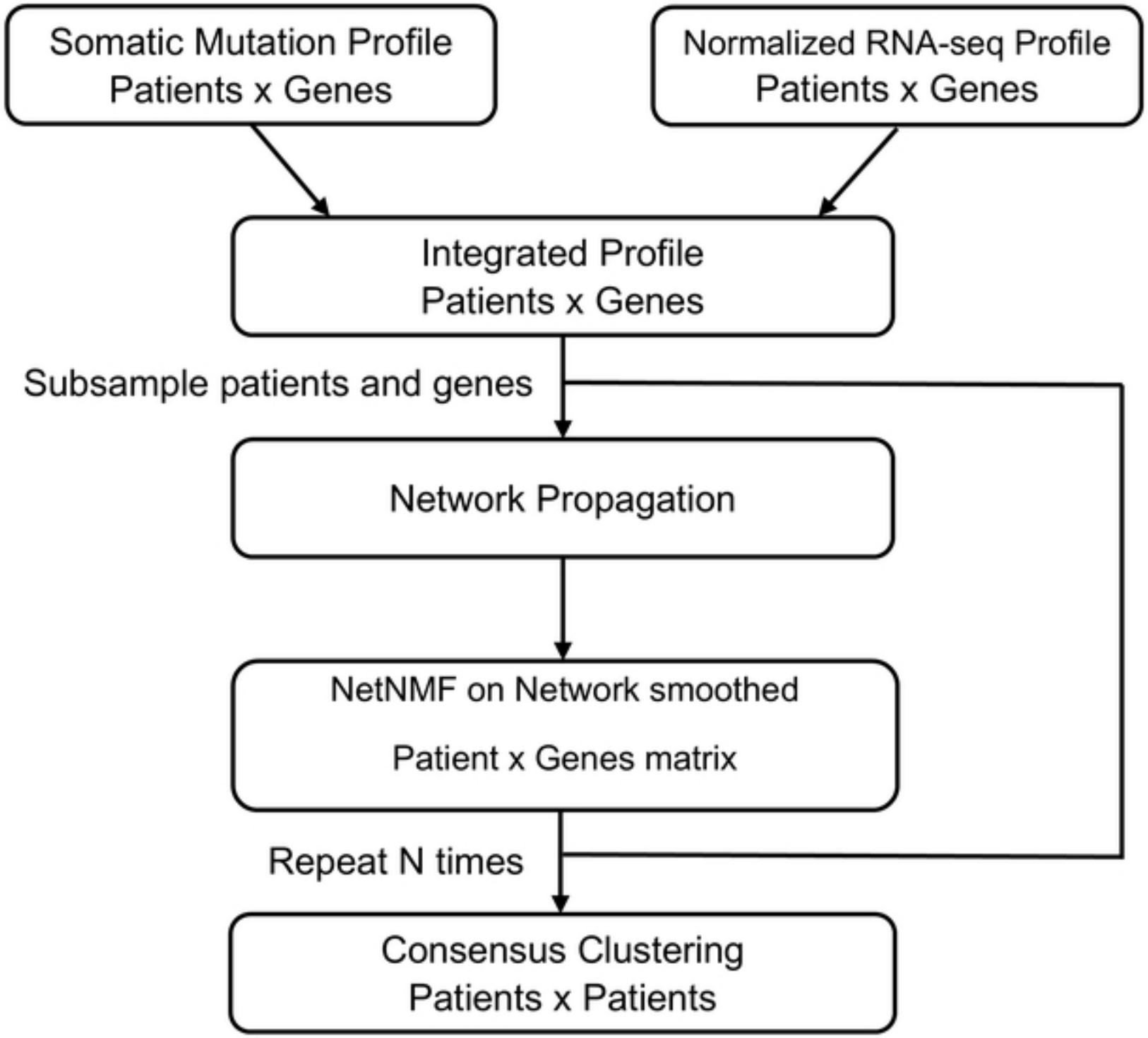
Somatic Mutation Profile
Patients x Genes
Normalized RNA-seq Profile
Patients x Genes
Integrated Profile
Patients x Genes
Subsample patients and genes
Network Propagation
NetNMF on Network smoothed
Patient x Genes matrix
Repeat N times
Consensus Clustering
Patients x Patients

Figure 1

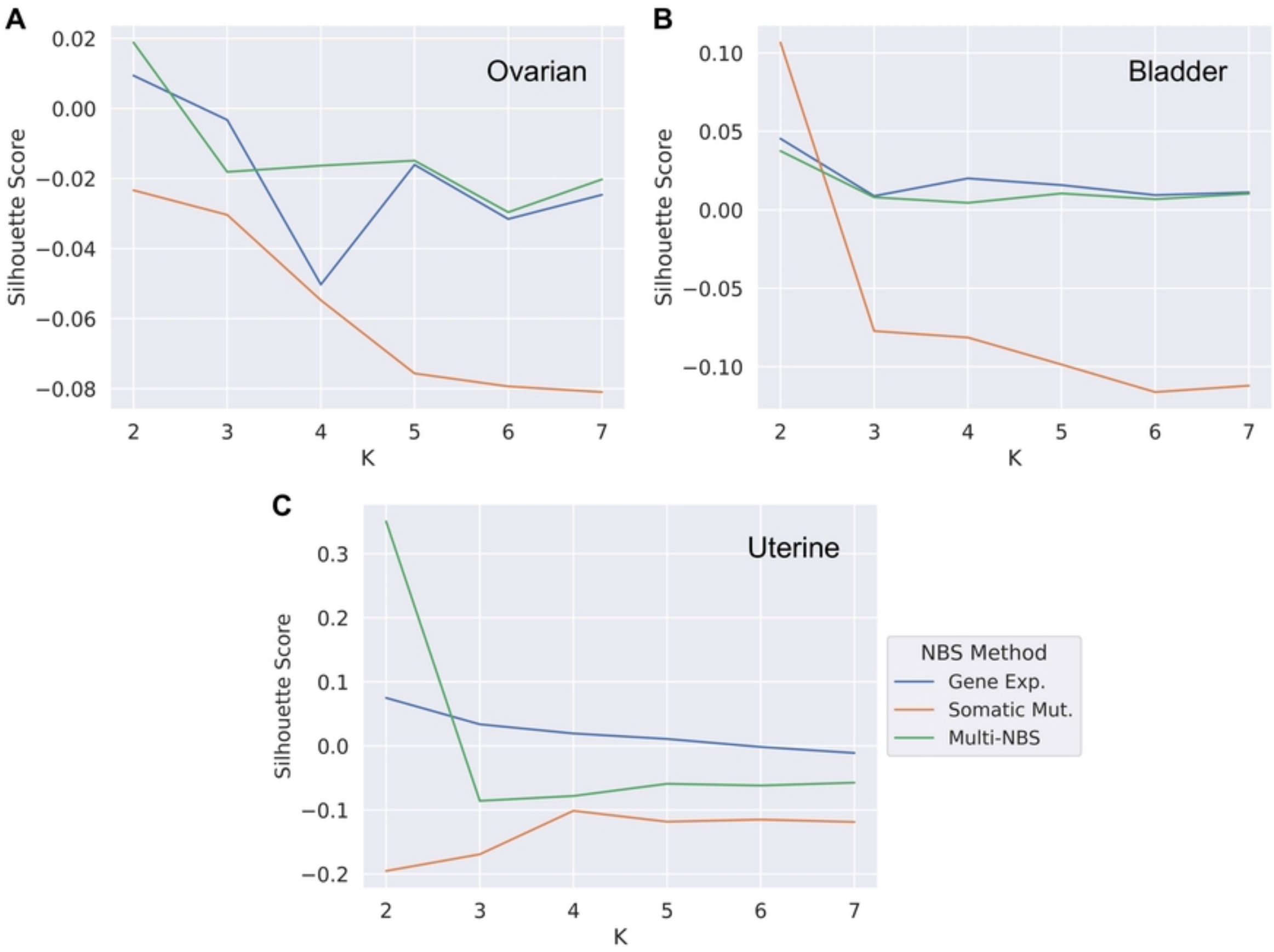
A
Ovarian
Silhouette Score
0.02
0.00
−0.02
−0.04
−0.06
−0.08
2
3
4
5
6
7
K
B
Bladder
Silhouette Score
0.10
0.05
0.00
−0.05
−0.10
2
3
4
5
6
7
K
C
Uterine
Silhouette Score
0.3
0.2
0.1
0.0
−0.1
−0.2
2
3
4
5
6
7
K
NBS Method
Gene Exp.
Somatic Mut.
Multi-NBS

Figure 2

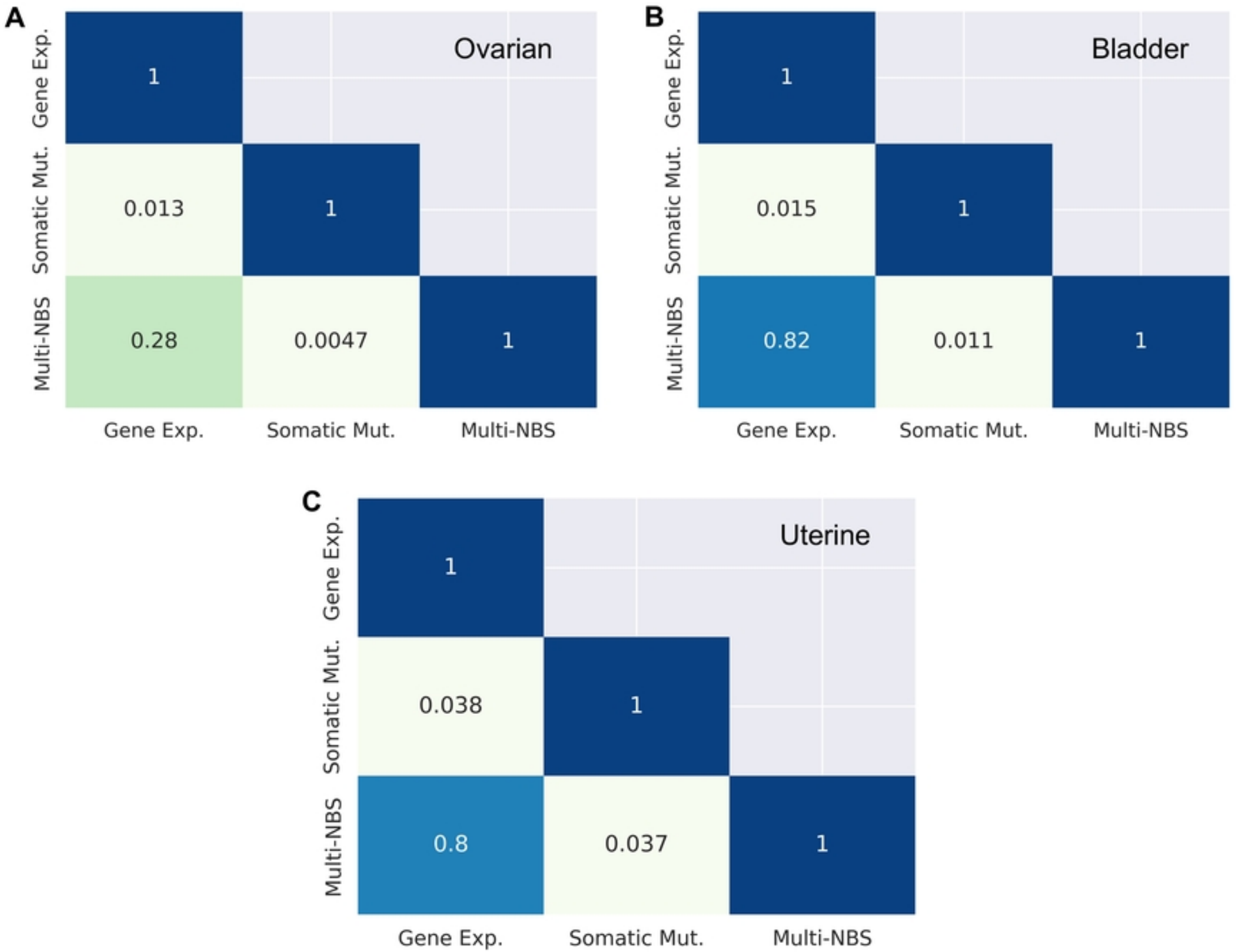
A
Ovarian
Gene Exp.
Somatic Mut.
Multi-NBS
1
0.013
1
0.28
0.0047
1
B
Bladder
1
0.015
1
0.82
0.011
1
C
Uterine
1
0.038
1
0.8
0.037
1

Figure 3

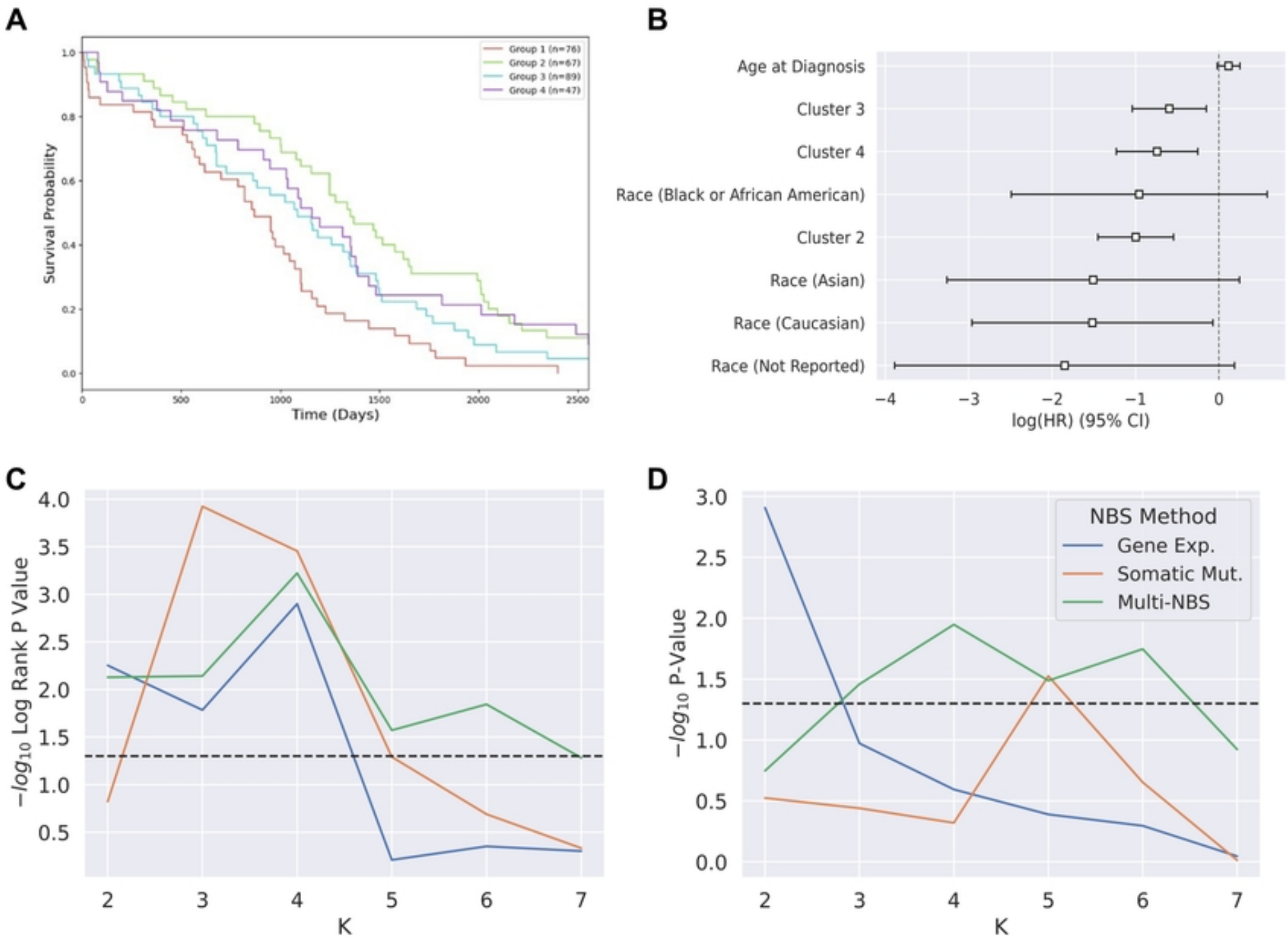
A
Group 1 (n=76)
Group 2 (n=67)
Group 3 (n=89)
Group 4 (n=47)
Survival Probability
Time (Days)
B
Age at Diagnosis
Cluster 3
Cluster 4
Race (Black or African American)
Cluster 2
Race (Asian)
Race (Caucasian)
Race (Not Reported)
log(HR) (95% CI)
C
$-log_{10}$ Log Rank P Value
K
D
NBS Method
Gene Exp.
Somatic Mut.
Multi-NBS
$-log_{10}$ P-Value
K

Figure 4

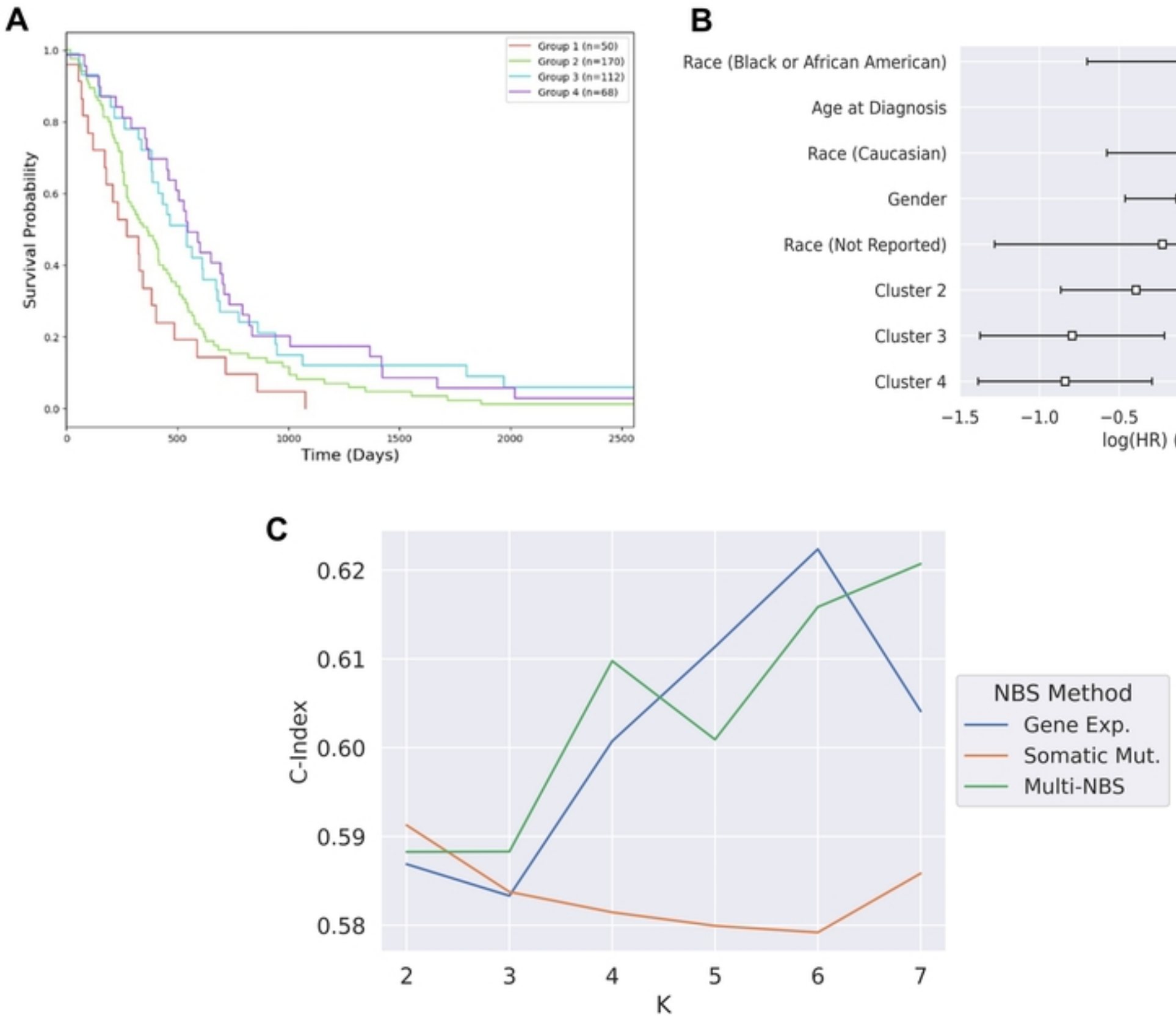
A
Group 1 (n=50)
Group 2 (n=170)
Group 3 (n=112)
Group 4 (n=68)
Survival Probability
Time (Days)
B
Race (Black or African American)
Age at Diagnosis
Race (Caucasian)
Gender
Race (Not Reported)
Cluster 2
Cluster 3
Cluster 4
log(HR) (95% CI)
C
C-Index
K
NBS Method
Gene Exp.
Somatic Mut.
Multi-NBS

Figure 5

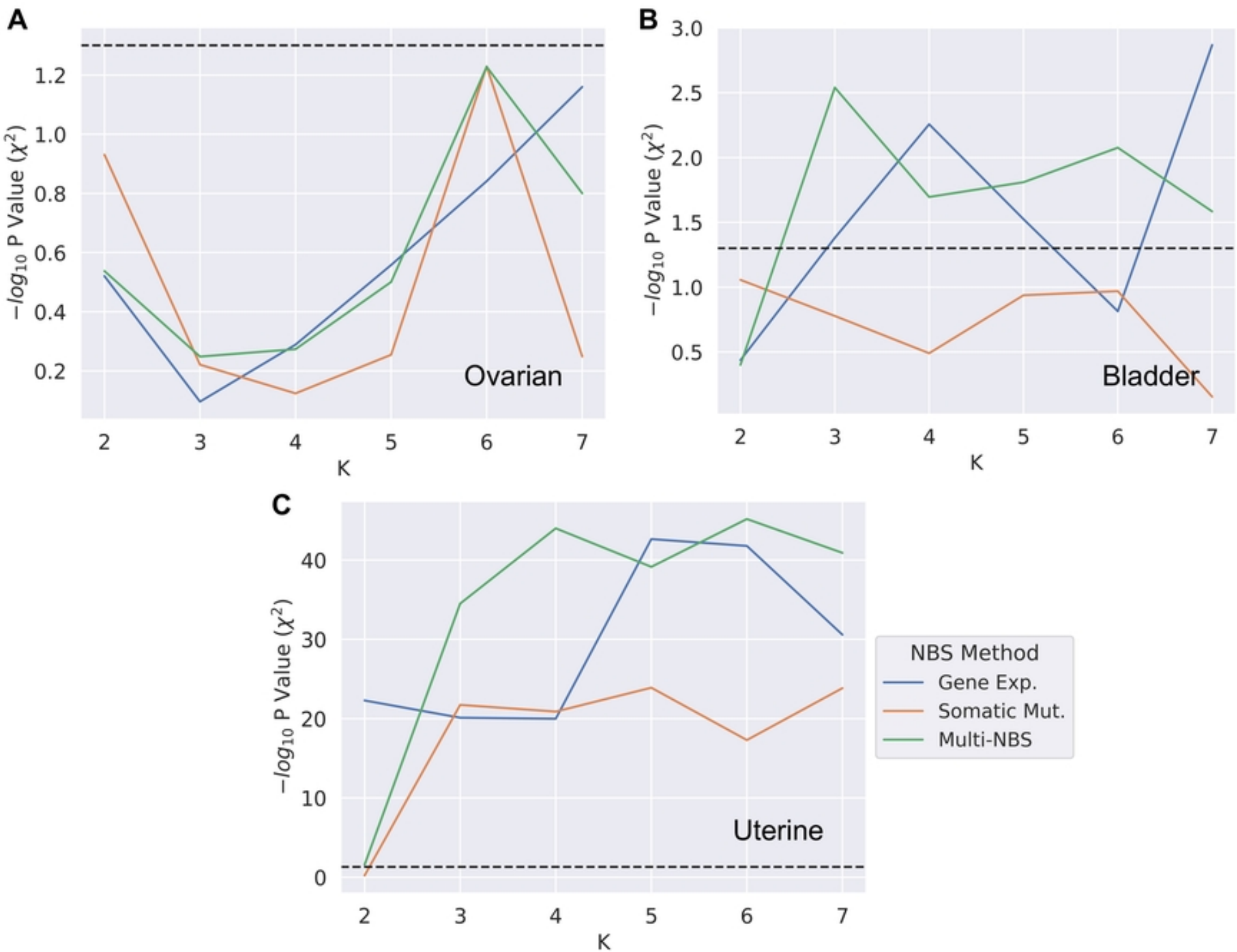
A
B
C
Ovarian
Bladder
Uterine
$-log_{10}$ P Value (χ^2)
K
NBS Method
Gene Exp.
Somatic Mut.
Multi-NBS

Figure 6