TITLE:

Cross-population analysis of high-grade serous ovarian cancer reveals only two robust subtypes

AUTHORS:

Gregory P. Way[a,b,c], James Rudd[a,d], Chen Wang[e], Habib Hamidi[f], Brooke L. Fridley[g], Gottfried Konecny[f], Ellen L. Goode[e], Casey S. Greene[a,b,h,1], Jennifer A. Doherty[a,d,2]

AFFILIATIONS:

[a]Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth College, Lebanon, NH; Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth College, Lebanon, NH

[b]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

[c]Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, PA 19103, USA

[d]Department of Epidemiology, Geisel School of Medicine at Dartmouth College, Lebanon, NH

[e]Department of Health Sciences Research, Mayo Clinic, Rochester, MN

[f]Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA

[g]Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS

[h]Department of Genetics, Geisel School of Medicine at Dartmouth College, Lebanon, NH

CO-CORRESPONDING AUTHOR:

[1]10-131 SCTR 34th and Civic Center Blvd, Philadelphia, PA 19104; Phone: 215-573-2991; Fax: 215-573-9135; csgreene@upenn.edu

[2]One Medical Center Drive, Lebanon, NH 03766; Phone: 603-653-9065; Fax: 603-653-9093; Jennifer.A.Doherty@Dartmouth.edu

AUTHOR EMAIL ADDRESSES

GW: gregway@upenn.edu
JR: james.e.rudd.gr@dartmouth.edu
CW: Wang.Chen@mayo.edu
HH: HHamidi@mednet.ucla.edu
BF: bfridley@kumc.edu
GK: GKonecny@mednet.ucla.edu
EG: egoode@mayo.edu
CG: csgreene@upenn.edu
JD: Jennifer.A.Doherty@dartmouth.edu

CONFLICTS OF INTEREST:

The authors do not declare any conflicts of interest.

OTHER PRESENTATIONS:

Aspects of this study were presented at the 2015 AACR Conference and the 2015 Rocky Mountain Bioinformatics Conference.

RUNNING HEAD:

Only two ovarian cancer subtypes are similar across populations

KEYWORDS:

Ovarian Cancer; Molecular Subtypes; Unsupervised Clustering; Reproducibility

NOTES:

Words: 3,392; Figures: 4; Tables 4; Sup. Figures: 9; Sup. Tables: 5; Sup. Methods

AUTHORS' CONTRIBUTIONS

Study concept and design: GW, JR, CG, JD. Original data collection and processing: CW, HH, BF, GK, EG. Data analysis: GW, JR, CG, JD. Manuscript drafting and editing: GW, JR, CG, JD. All authors read, commented on, and approved the final manuscript.

ABSTRACT:

Background

Four gene expression-based subtypes of high-grade serous ovarian cancer (HGSC), variably associated with differential survival, have been previously described. However, in these studies, clustering heuristics were consistent with only three subtypes and reproducibility of the subtypes across populations and assay platforms has not been formally assessed. Therefore, we systematically determined the concordance of transcriptomic HGSC subtypes across populations.

Methods

We used a unified bioinformatics pipeline to independently cluster (k = 3 and k = 4) five mRNA expression datasets with >130 tumors using k-means and non-negative matrix factorization (NMF) without removing "hard-to-classify" samples. Within each population, we summarized differential expression patterns for each cluster as moderated t statistic vectors using Significance Analysis of Microarrays. We calculated Pearson's correlations of these vectors to determine similarities and differences in expression patterns between clusters. We identified sets of clusters

that were most correlated across populations to define syn-clusters (SC), and we associated SC expression patterns with biological pathways using geneset overrepresentation analyses.

Results

Across populations, for $k = 3$, moderated t score correlations for clusters 1, 2 and 3 ranged between 0.77-0.85, 0.80-0.90, and 0.65-0.77, respectively. For $k = 4$, correlations for clusters 1-4 were 0.77-0.85, 0.83-0.89, 0.51-0.76, and 0.61-0.75, respectively. Within populations, comparing analogous clusters ($k = 3$ versus $k = 4$), correlations were high for clusters 1 and 2 (0.91-1.00), but lower for cluster 3 (0.22-0.80). Results were similar using NMF. SC1 corresponds to mesenchymal-like, SC2 to proliferative-like, SC3 to immunoreactive-like, and SC4 to differentiated-like subtypes reported previously.

Conclusions

While previous single-population studies reported four HGSC subtypes, our cross-population comparison finds strong evidence for only two subtypes and our re-analysis of previous data suggests that results favoring four subtypes may have been driven, at least in part, by the inclusion of samples with low malignant potential. Because the mesenchymal-like and proliferative-like subtypes are highly consistent across populations, they likely reflect intrinsic biological subtypes and are strong candidates for targeted therapies. The other two previously described subtypes (immunoreactive-like and differentiated-like) are considerably less consistent and may represent either a single subtype or signal that is not amenable to clustering.

INTRODUCTION:

Invasive ovarian cancer is a heterogeneous disease typically diagnosed at a late stage, with high mortality [1] . The most aggressive and common histologic type is high-grade serous (HGSC) [2], characterized by extensive copy number variation and TP53 mutation [3]. Given the

genomic complexity of these tumors, mRNA expression can be thought of as a summary measure of these genomic and epigenetic alterations, to the extent that the alterations influence gene expression. Efforts to use whole genome mRNA expression analyses to stratify HGSC into clinically relevant subtypes have yielded potentially promising results, with all studies to date observing three to four subtypes with varying components of mesenchymal, proliferative, immunoreactive, and differentiated gene expression signatures [3–6], and some studies observing survival differences across subtypes [4, 5] . Tothill *et al*. first identified four HGSC subtypes (as well as two other subtypes which largely included low grade and low malignant potential samples) in an Australian population using *k*-means clustering. The authors labeled the subtypes as C1-C6, and observed that women with the C1 subtype, with a stromal-like gene signature, experienced the poorest survival compared to the other subtypes [4] . Later, The Cancer Genome Atlas (TCGA), in an assemblage of tumors from various institutions throughout The United States, used non-negative matrix factorization (NMF) clustering and also reported four subtypes which they labeled as ‘mesenchymal’, ‘differentiated’, ‘proliferative’, and ‘immunoreactive’, but there were no observed differences in survival [3] . The TCGA group also applied NMF clustering to the Tothill data, and noted that analogous subtypes had similar significantly differentially expressed genes [3]. Konecny *et al.* also applied NMF to cluster HGSC samples from the Mayo Clinic and reported four subtypes, which they labeled as C1-C4 [5] . While these subtypes are similar to those described by TCGA, the Konecny *et al*. refined classifier was better able to differentiate survival between groups in their own data, and in data from TCGA and Bonome *et al.* [6] . In the Konecny *et al.* population, as similarly observed in Tothill *et al*., the mesenchymal-like (described as stromal-like in Tothill *et al.*) and proliferative-like subtypes had poor survival, and the immunoreactive-like subtype had favorable survival [5] .

While results from these studies are relatively consistent, there exist several limitations inherent to each study that must be overcome in order to confirm the existence and reproducibility of HGSC subtypes across different populations. For example, in more recent TCGA analyses by the Broad Institute Genome Data Analysis Center (GDAC) Firehose initiative with the largest number of HGSC cases evaluated to date (n = 569), three subtypes fit the data better than did four [7, 8]. Also, in the original analysis of the TCGA data, over 80% of the samples were assigned to more than one subtype [9], as were 42% of the Mayo samples in Konecny *et al*. Furthermore, in both TCGA and Tothill *et al.*, ~8-15% of samples were not able to be classified and were excluded. Because of this uncertainty in HGSC subtyping, further characterization is essential in order to determine whether homogeneous, unique subtypes exist and to subsequently identify etiologic factors and to develop targeted treatments.

To comprehensively characterize subtypes, we analyzed data from five independent populations using a unified bioinformatics pipeline. Previous approaches to integrate subtype analyses across populations either removed samples that were difficult to classify or identified subtypes in a single population, built a classifier for those subtypes within the population, and applied the classifier to other populations to label samples [3, 5]. In contrast with these population-specific approaches, our pipeline performs unsupervised clustering using both *k*-means clustering and NMF separately in each population without removing "hard-to-classify" samples. This allows us to systematically identify and compare the presence and reproducibility of subtypes in the largest study of HGSC subtypes to date. We summarize the expression patterns of over 10,000 genes for each identified subtype and comprehensively characterize correlations between subtype-specific gene expression both within and between populations. We

identify reproducible clusters characterized by patterns of differentially expressed genes that are positively correlated across populations, which we term "syn-clusters" (SC).

METHODS:

Data Inclusion

We applied inclusion criteria as described in the supplementary materials using data from the R package, curatedOvarianData [10] (Table S1) and a separate dataset ("Mayo") [5]. We deposited the Mayo HGSC samples as well as other samples with mixed histologies and grades, for a total of 528 additional ovarian tumor samples, in NCBI's Gene Expression Omnibus (GEO) [11]. The data can be accessed with the accession number GSE74357 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74357). All uploaded tumor samples were collected with approval by an institutional review board and by the U.S. Department of Health and Human Services. After applying the unified inclusion criteria pipeline, our final analytic datasets include: TCGA (n = 499) [3, 7]; Mayo (n = 379; GSE74357) [5]; Yoshihara (n = 256; GSE32062.GPL6480) [12]; Tothill (n = 241; GSE9891) [4]; and Bonome (n = 185; GSE26712) [6] (Table 1). We analyzed the intersection of genes measured in all five populations, which included 10,930 genes (Supplementary Fig. S1). Software to replicate all analyses is provided under a permissive open source license [13].

Clustering

Because 3 or 4 subtypes had been reported previously, and the heuristics for the TCGA firehose analyses, the original HGSC TCGA publication, and Konecny *et al.* suggest 3 subtypes, we focused on examining cluster assignment within and across populations for clusters identified

using $k = 3$ or $k = 4$. As detailed in the supplemental methods, we identified the 1,500 genes with the highest variance from each population and used the union of these genes (n = 3,698) for subsequent analyses. We performed *k*-means clustering on these 3,698 genes in each population using the R package "cluster" (version 2.0.1) [14] with 20 initializations, and we characterized patterns of sample assignment changes when $k = 3$ versus $k = 4$. We further characterized clustering solutions within populations using sample-by-sample Pearson's correlation matrices. We repeated these analyses using NMF in the R package "NMF" (version 0.20.5) [15] with 100 initializations used for each *k*. Lastly, we investigated the reproducibility of previous HGSC subtyping studies identifying four subtypes by calculating cophenetic correlation coefficients following NMF with 10 consensus runs for $k = 2$ through 8.

Identification of Syn-Clusters

We performed a significance analysis of microarray (SAM) [16, 17] analysis on all clusters from each population for $k = 3$ and $k = 4$ using all 10,930 genes. This resulted in a cluster-specific moderated *t* statistic for each of the input genes [18]. To summarize the expression patterns of all 10,930 genes for a specific cluster in a specific population, we combined each gene's moderated *t* statistics for one versus all comparisons into a vector of length 10,930. To generate comparable labels across $k = 3$ and $k = 4$ analyses, the $k = 3$ cluster which was most strongly correlated with a $k = 4$ cluster in the TCGA data was labeled "cluster 1" and the second strongest "cluster 2" etc. Clusters in other populations that were most strongly correlated with the TCGA clusters were assigned the same label. For each centroid assignment, clusters most positively correlated across populations form a syn-cluster (SC); i.e. the clusters from each population that are most correlated with each other and with TCGA "cluster 1" belong

to SC1. We also compared our sample assignments to subtypes reported in the Tothill, TCGA, and Konecny publications.

Identifying Biological Processes Associated with Syn-Clusters

To annotate the SCs with associated biological processes, we first identified the statistically significantly differentially expressed genes in each SAM list. We used a Bonferroni adjustment taking into account the total number of genes considered (10,930) resulting in a *p*-value cutoff of 4.6×10^{-6}. We used the intersection of these cluster-specific genesets across populations to create the final SC associated genesets. We then input these SC associated genesets into a PANTHER analysis [19] to determine SC-specific overrepresented biological pathways (Supplementary Materials).

RESULTS:

Sample Cluster Assignment

To visually inspect the consistency and distinctness of clusters, we compared sample-by-sample correlation heatmaps (Fig. 1). For both *k* values and in each population, we observed high sample-by-sample correlations within clusters and relatively low sample-by-sample correlations across clusters (Fig. 1). The clusters in the Bonome population are depicted in gray scale because in cross-population analyses to identify SCs their expression patterns did not correlate with the clusters observed consistently in the four other populations (Table 2). Clustering results using NMF are similar to the *k* means assignments (Supplementary Fig S2.)

To better understand the changes in cluster assignment for $k = 3$ versus $k = 4$, we compared the number of samples belonging to each cluster within each population (excluding

Bonome; Fig. 2). Overall, the cross-*k* pattern was consistent across populations. Cluster 1 contained essentially the same samples for both $k = 3$ and $k = 4$, as did cluster 2, but samples from cluster 3 when $k = 3$ tended to be split between clusters 3 and 4 when $k = 4$. Additionally, cluster 3 in $k = 4$ tended to have varying numbers of samples from cluster 1 in $k = 3$, and cluster 4 in $k = 4$ tended to include some samples from cluster 2 in $k = 3$ (Fig. 2). These specific patterns were consistent in each population.

Correlation of Cluster-Specific Expression Patterns

Within populations, we observed high Pearson correlations of moderated *t* score vectors between analogous clusters across $k = 3$ and $k = 4$ (Table 2). Across populations, we observed strong positive correlations of moderated *t* score vectors between analogous clusters in the TCGA, Tothill, Mayo, and Yoshihara cluster assignments (Fig. 3; Table 3). However, while the clusters across $k = 3$ and $k = 4$ were correlated within the Bonome data, they did not correlate strongly with clusters identified in the other populations (Table 3). Because the correlations were low compared to those observed in all four other populations, the Bonome data were not included in subsequent analyses. Across populations, positive correlations between clusters belonging to the same SC, and negative correlations between clusters in different SCs, were stronger for clusters identified when $k = 3$ than when $k = 4$ (Figure 3). This analysis effectively demonstrates that three subtypes fit the data more consistently than do the previously accepted four subtypes. We observed strong positive correlations for both SC1 and SC2 within and across all populations. Importantly, we also observed strong negative correlations between SC1 and SC2 within and between populations (ranging from -0.42 to -0.63 for $k = 3$ and -0.12 to -0.64 for $k = 4$). Weaker and more variable positive correlations were observed for SC3 and SC4 across

populations. For k = 4, Yoshihara cluster 3 appears to be correlated to both clusters 3 and 4 in the other populations, and cluster 4 to be weakly correlated to cluster 2 in the other populations. In contrast, for k = 3, SC3 is positively correlated across populations (though more weakly than SC1 and SC2 correlations across populations), and tends to be uncorrelated or inversely correlated with SC1, and consistently inversely correlated with SC2.

Within each population, clusters identified by NMF were very similar to those identified using *k*-means clustering (Fig. 4). Again, both positive and negative correlations are stronger for k = 3 than for k = 4. Across k = 3 and k = 4, correlations are strongest for clusters 1 and 2. Sample cluster assignments for both *k*-means and NMF clusters are provided in Table S2.

Reproducibility of previous HGSC subtyping studies

We evaluated the number of subtypes that fit the data best by observing the cophenetic correlation coefficients within individual datasets, setting NMF to find 2 through 8 clusters inclusively. We observed a similar pattern in each population (Supplementary Fig S3 – S7) in which the highest cophenetic correlation was reached for two clusters and, based on the heatmaps, appeared to have the highest consensus. Importantly, in each dataset, four clusters were not observed to represent the data better than two or three.

Comparison with previously-identified HGSC clusters

Our clustering results for the Tothill, TCGA, and Mayo datasets are highly concordant with the clustering described in the original publications [3–5], as evidenced by the high degree of overlap in sample assignments to the previously-defined clusters (Table 4). Our SC1 for both *k*-means analyses was mapped to the “Mesenchymal” label from TCGA, “C1” from Tothill, and

mostly to "C4" from Mayo. SC1 was the most stable in our analysis within all datasets, across *k* = 3 and *k* = 4, and across clustering algorithms. SC2, which was also observed consistently, was most similar to the "Proliferative" label from TCGA, "C5" from Tothill, and "C3" from Mayo. SC3 for *k* = 3 was associated with both the "Immunoreactive" and "Differentiated" TCGA labels, "C2" and "C4" in Tothill, and "C1" and "C2" in Mayo. When setting *k*-means to find four clusters, SC3 was associated with "Immunoreactive", "C2", and "C1" while SC4 was associated with "Differentiated", "C4", and "C2" for TCGA, Tothill, and Mayo respectively. Pathway analysis results for all SCs are summarized in more detail in the supplementary materials and are presented in supplementary table S5.

DISCUSSION:

Single-population studies have reported four subtypes of HGSC [3–5, 7], but it is difficult to compare the results because each study performed analyses with different sample inclusion criteria and different statistical methods. To address this, we used uniform sample inclusion criteria and applied *k*-means clustering and NMF through a standardized bioinformatics pipeline to five independent HGSC datasets including American, Australian, and Japanese women to systematically characterize HGSC subtypes within and between populations.

This allowed us to identify syn-clusters (SC), or groups of analogous clusters observed across populations. Despite considerable diversity in the populations studied and the assay platforms used, in four of the five populations studied, we identified two distinct and highly robust SCs (SC1 and SC2). SC1 and SC2 consisted of mostly the same samples when *k* = 3 and *k* = 4, and global differential gene expression patterns were similar in across populations. Taken together, these observations increase our confidence that each of these clusters represents a set of

reproducible biological signals. The strong positive correlations of subtype-specific gene expression signatures indicate homogeneity of gene expression patterns across populations for SC1 and SC2. The strong negative correlations between SC1 and SC2 also indicate that they are distinct from one another; this is emphasized by the inverse direction of expression for the immune system process genes.

We also identified a third SC (SC3) and potentially fourth SC (SC4), though both positive correlations across populations and negative correlations between these and other subtypes within populations are weaker. In fact, we observed a positive correlation between clusters that belong to these SCs within some populations, particularly in the Japanese population. In contrast with the previous reports of four subtypes of HGSC [3–5], we observed that both concordance of analogous subtypes and inverse correlations between distinct subtypes were stronger for analyses of three clusters as opposed to four.

Because cross-population comparisons suggest that three clusters show more consistency than four, we explored within-study heuristics that suggested four subtypes in previous research. Heuristics using our sample inclusion criteria are consistent with two or three as opposed to four subtypes in each population, which is in fact consistent with results from analyses performed in the previous reports. The cophenetic coefficient measures how precisely a dendrogram retains sample by sample pairwise distances and can be used to compare clustering accuracy [20]. While both Konecny and TCGA reported four subtypes, in both analyses $k = 3$ resulted in a higher cophenetic coefficient than $k = 4$ and larger drops in cophenetic coefficient were observed after three than four, indicating that 3 subtypes fit the data better than 4 [3], [5]. However, in TCGA's re-analysis of the Tothill *et al*. HGSC samples shown in supplemental Figure S6.2 in their publication, the cophenetic coefficient dropped dramatically at $k = 3$ before recovering at $k = 4$,

which would support the existence of 4 subtypes [3]. Notably, TCGA's figure legend for this supplemental result indicates that they did not remove samples with low malignant potential (LMP) from the Tothill data. Our analysis of Tothill *et al.* differed from TCGA's in that our unified pipeline specifically removed LMP samples, and instead supports the existence of 2 or 3 subtypes (Supplemental Figure 6). To evaluate the influence of the LMP samples in the Tothill data, we repeated our analyses including them, and observed a drop in the cophenetic coefficient for $k = 3$ relative to $k = 4$ (Supplementary Figure 8). This suggests that the 4 subtypes observed in TCGA's analysis of the Tothill data may be due, in part, to the inclusion of LMP samples.

In our study, results for each population were similar for the *k* means and NMF clustering algorithms that were applied in previous studies, further validating both the presence of two reproducible subtypes and the less stable nature of the other subtypes. Compared to the clusters reported in TCGA, Tothill, and Konecny, SC1 was most similar to the mesenchymal/C1/C4 subtype and SC2 was most similar to the proliferative/C5/C3 subtype, respectively. While concordance between the original Tothill and TCGA subtypes was reported in the TCGA HGSC publication [3], our analysis included an additional 59 TCGA samples. As well, we included an additional 210 samples from Mayo that were not analyzed in the original Konecny *et al.* publication [5]. We did not observe strong patterns in survival differences across the subtypes that we identified (see Supplementary Material). However, we would not necessarily expect to find differences in survival unless the biological characteristics of the tumor subtypes translate into different responses to standard treatments. Instead, the goal is to identify the most consistent subtypes so that they can be exhaustively characterized and targeted treatments can be developed [21].

The consistency of SC1 and SC2 across *k* parameters and between diverse populations is remarkable for a number of reasons. While these studies represent the largest collections of HGSC tumors to date, given the difficulties in collecting fresh frozen tissue for large-scale gene expression studies, it is unclear how accurately any of these data sets reflect the underlying population distribution of HGSC subtypes. Results from gene expression/RNA sequencing assays in large, population-based formalin-fixed paraffin-embedded (FFPE) tumor collections will be important in further informing the definitions of HGSC subtypes. Given the intra-tumor heterogeneity that is likely to exist [22], our approach would be strengthened by having data on multiple areas of the tumors. Finally, since histology and grade classification have changed over time [23, 24], it is unclear whether the populations we studied used comparable guidelines to determine histology and grade. We attempted to exclude all low grade serous and endometrioid samples because they often have very different gene expression patterns and more favorable survival compared to their higher grade counterparts [2]. While the Bonome publication specified that they included only high-grade tumors, grade is not included in the Bonome GSE26712 data set, so we were unable to determine whether the grade distribution differs from the other studies [6] . At any rate, it is unclear why the Bonome clusters, while internally consistent across *k*, did not correspond to the clusters observed in other populations. If samples are misclassified with respect to grade or other characteristics, depending on the extent of the misclassification, lower correlations and consequently difficulty assigning SCs could result.

In summary, our study demonstrates that two SCs of HGSC, “mesenchymal-like” and “proliferative-like”, are clearly and consistently identified within and between populations. This suggests that there are two reproducible HGSC subtypes that are either etiologically distinct, or acquire phenotypically determinant alterations through their development. These two SCs have

different sets of significantly enriched pathways, which may indicate distinct processes regulating tumor progression. The “mesenchymal-like” subtype includes genes involved with extracellular matrix and cell to cell adhesion processes, while the “proliferative-like” subtype includes lower expressed immune-related genes, consistent with previous studies which have identified a negative immune signature in this subtype [5]. Our study also suggests that the previously described “immunoreactive-like” and “differentiated-like” subtypes appear more variable across populations. These may represent, for example, steps along an immunoreactive continuum or could represent the basis of a third, but more variable subtype. Because the “mesenchymal-like” and “proliferative-like” subtypes are consistently observed within and between populations, at the current time these subtypes are the strongest candidates for development of subtype-specific treatment strategies.

ACKNOWLEDGEMENTS:

We would like to thank Sebastian Armasu and Hsiao-Wang Chen for help with statistical analyses and data processing and Emily Kate Shea for helpful discussions.

FUNDING:

This work was supported by the Institute for Quantitative Biomedical Sciences; the Norris Cotton Cancer Center Developmental Funds; the National Cancer Institute at the National Institutes of Health (R01 CA168758 to J.A.D., F31 CA186625 to J.R., R01 CA122443 to E.L.G.); the Mayo Clinic Ovarian Cancer SPORE (P50 CA136393 to E.L.G.); the Mayo Clinic Comprehensive Cancer Center-Gene Analysis Shared Resource (P30 CA15083); the Gordon and

Betty Moore Foundation's Data-Driven Discovery Initiative (grant number GBMF 4552 to C.S.G.); and the American Cancer Society (grant number IRG 8200327 to C.S.G.).

FIGURE LEGENDS:

Figure 1. Sample by sample Pearson correlation matrices. Top panel: $k = 3$. Bottom panel: $k = 4$. The color bars are coded as blue, syn-cluster 1 (SC1); red, SC2; green, SC3; and purple, SC4. In the matrices, red represents high correlation, blue low correlation, and white intermediate correlation. The scales are slightly different in each population because of different correlational structures. The grey Bonome clusters indicate clusters not correlating well with any cluster from the other populations.

Figure 2. Sample membership distribution changes when setting k means to find $k = 3$ and $k = 4$. The bars represent sample cluster membership with $k = 4$ and the colors indicate the same samples' cluster assignments for when $k = 3$. Samples from the third cluster with $k = 3$ tend to split apart to form the third and fourth clusters when $k = 4$.

Figure 3. SAM moderated t score Pearson correlations reveal consistency across populations. The color bars are coded as blue, syn-cluster 1 (SC1); red, SC2; green, SC3; and purple, SC4. (A) Correlations across datasets for *k* means $k = 3$. (B) Correlations across datasets for k means $k = 4$. The matrices are symmetrical and the upper triangle holds scatter plots for each comparison where each point represents one of the 10,930 genes measured in each population. For $k = 3$ and $k = 4$, clusters associated with SC1 and SC2 are highly consistent across populations.

Figure 4. SAM moderated t score Pearson correlations of clusters formed by *k* means clustering and NMF clustering reveals consistency between clustering methods. Results are shown for both methods when setting each algorithm to find 3 and 4 clusters. The color bars are coded as blue, syn-cluster 1 (SC1); red, SC2; green, SC3; and purple, SC4.

Supplementary Figure S1. Overlapping genes assayed using either the HG-U1133 Affymetrix platform (TCGA, Tothill, Bonome) or the Agilent 4x44K platform (Mayo, Yoshihara). Differences across datasets arise from inherent array differences and/or differences in quality control preprocessing.

Supplementary Figure S2. NMF consensus matrices for datasets when (A) $k = 3$ and (B) $k = 4$. The first track represents cluster membership for k means clusters and the second track represents silhouette widths. Note however that the NMF clusters are not mapped to the ordered *k* means clusters.

Supplementary Figure S3. TCGA dataset (n = 499). Consensus NMF clustering of 3,698 most variably expressed genes across five HGSC ovarian cancer datasets. Data displays consensus clustering for $k = 2$ to $k = 6$ for 10 NMF runs alongside the cophenetic correlation results for $k = 2$ to $k = 8$.

Supplementary Figure S4. Mayo dataset (n = 379). Consensus NMF clustering of 3,698 most variably expressed genes across five ovarian cancer datasets (all high grade serous samples were

419 retained). Data displays consensus clustering for k = 2 to k = 6 for 10 NMF runs alongside the
420 cophenetic correlation results for k = 2 to k = 8.
421
422 **Supplementary Figure S5.** Yoshihara dataset (n = 256). Consensus NMF clustering of 3,698
423 most variably expressed genes across five HGSC ovarian cancer datasets. Data displays
424 consensus clustering for k = 2 to k = 6 for 10 NMF runs alongside the cophenetic correlation
425 results for k = 2 to k = 8.
426
427 **Supplementary Figure S6.** Tothill dataset (n = 242). Consensus NMF clustering of 3,698 most
428 variably expressed genes across five HGSC ovarian cancer datasets. Data displays consensus
429 clustering for k = 2 to k = 6 for 10 NMF runs alongside the cophenetic correlation results for k =
430 2 to k = 8.
431
432 **Supplementary Figure S7.** Bonome dataset (n = 185). Consensus NMF clustering of 3,698
433 most variably expressed genes across five HGSC ovarian cancer datasets. Data displays
434 consensus clustering for k = 2 to k = 6 for 10 NMF runs alongside the cophenetic correlation
435 results for k = 2 to k = 8.
436
437 **Supplementary Figure S8.** Tothill dataset including low malignant potential samples (n = 260).
438 Consensus NMF clustering of 3,698 most variably expressed genes across five HGSC ovarian
439 cancer datasets. Low malignant potential (borderline) samples (n = 18) were not removed prior
440 to clustering (these samples were removed in Supplementary Figure S6). Data displays

consensus clustering for k = 2 to k = 6 for 10 NMF runs alongside the cophenetic correlation results for k = 2 to k = 8.

Supplementary Figure S9. Kaplan-Meier survival curves. For each population, the top plot is for *k* = 3 and the bottom plot is for *k* = 4.

REFERENCES:

1. 1. Kurman RJ, Shih I-M: **The Origin and Pathogenesis of Epithelial Ovarian Cancer: A Proposed Unifying Theory:** *Am J Surg Pathol* 2010, **34**:433–443.

2. Vang R, Shih I-M, Kurman RJ: **Ovarian Low-grade and High-grade Serous Carcinoma: Pathogenesis, Clinicopathologic and Molecular Biologic Features, and Diagnostic Problems**. *Adv Anat Pathol* 2009, **16**:267–282.

3. The Cancer Genome Atlas: **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011, **474**:609–615.

4. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, Johnson DS, Trivett MK, Etemadmoghadam D, Locandro B, Traficante N, Fereday S, Hung JA, Chiew Y-E, Haviv I, Australian Ovarian Cancer Study Group, Gertig D, deFazio A, Bowtell DDL: **Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome**. *Clin Cancer Res* 2008, **14**:5198–5208.

5. Konecny GE, Wang C, Hamidi H, Winterhoff B, Kalli KR, Dering J, Ginther C, Chen H-W, Dowdy S, Cliby W, Gostout B, Podratz KC, Keeney G, Wang H-J, Hartmann LC, Slamon DJ, Goode EL: **Prognostic and Therapeutic Relevance of Molecular Subtypes in High-Grade Serous Ovarian Cancer**. *JNCI J Natl Cancer Inst* 2014, **106**:dju249–dju249.

6. Bonome T, Levine DA, Shih J, Randonovich M, Pise-Masison CA, Bogomolniy F, Ozbun L, Brady J, Barrett JC, Boyd J, Birrer MJ: **A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer**. *Cancer Res* 2008, **68**:5478–5486.

7. Broad Institute TCGA Genome Data Analysis Center: **Analysis Overview for Ovarian Serous Cystadenocarcinoma (Primary solid tumor cohort) - 15 July 2014**. 2014.

8. Broad Institute TCGA Genome Data Analysis Center: **Clustering of mRNA expression: consensus NMF**. 2015.

9. Verhaak RGW, Tamayo P, Yang J-Y, Hubbard D, Zhang H, Creighton CJ, Fereday S, Lawrence M, Carter SL, Mermel CH, Kostic AD, Etemadmoghadam D, Saksena G, Cibulskis K,

Duraisamy S, Levanon K, Sougnez C, Tsherniak A, Gomez S, Onofrio R, Gabriel S, Chin L, Zhang N, Spellman PT, Zhang Y, Akbani R, Hoadley KA, Kahn A, Köbel M, Huntsman D, et al.: **Prognostically relevant gene signatures of high-grade serous ovarian carcinoma**. *J Clin Invest* 2012.

10. Ganzfried BF, Riester M, Haibe-Kains B, Risch T, Tyekucheva S, Jazic I, Wang XV, Ahmadifar M, Birrer MJ, Parmigiani G, Huttenhower C, Waldron L: **curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome**. *Database* 2013, **2013**:bat013–bat013.

11. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30**:207–210.

12. Yoshihara K, Tsunoda T, Shigemizu D, Fujiwara H, Hatae M, Fujiwara H, Masuzaki H, Katabuchi H, Kawakami Y, Okamoto A, Nogawa T, Matsumura N, Udagawa Y, Saito T, Itamochi H, Takano M, Miyagi E, Sudo T, Ushijima K, Iwase H, Seki H, Terao Y, Enomoto T, Mikami M, Akazawa K, Tsuda H, Moriya T, Tajima A, Inoue I, Tanaka K, et al.: **High-Risk Ovarian Cancer Based on 126-Gene Expression Signature Is Uniquely Characterized by Downregulation of Antigen Presentation Pathway**. *Clin Cancer Res* 2012, **18**:1374–1385.

13. Gregory Way, James Rudd, Casey Greene: **Analytical Code for "Cross-population analysis of high-grade serous ovarian cancer reveals only two robust subtypes."** 2015.

14. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K: **cluster: Cluster Analysis Basics and Extensions**. 2014, **R package version 1.15.3.**

15. Brunet J-P, Tamayo P, Golub TR, Mesirov JP: **Metagenes and molecular pattern discovery using matrix factorization**. *Proc Natl Acad Sci* 2004, **101**:4164–4169.

16. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci* 2001, **98**:5116–5121.

17. Schwender H, Krause A, Ickstadt K: **Identifying interesting genes with sigenes**. *RNews* 2006, **6**:45–50.

18. Schwender H: **siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches.** 2012, **R package version 1.40.0**.

19. Mi H, Muruganujan A, Thomas PD: **PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees**. *Nucleic Acids Res* 2013, **41**:D377–D386.

20. Sokal RR, Rohlf FJ: **The Comparison of Dendrograms by Objective Methods**. *Taxon* 1962, **11**:33.

21. Waldron L, Riester M, Birrer M: **Molecular Subtypes of High-Grade Serous Ovarian Cancer: The Holy Grail?** *JNCI J Natl Cancer Inst* 2014, **106**:dju297–dju297.

22. Blagden SP: **Harnessing Pandemonium: The Clinical Implications of Tumor Heterogeneity in Ovarian Cancer**. *Front Oncol* 2015, **5**.

23. Silverberg SG: **Histopathologic grading of ovarian carcinoma: a review and proposal**. *Int J Gynecol Pathol Off J Int Soc Gynecol Pathol* 2000, **19**:7–15.

24. Soslow RA: **Histologic Subtypes of Ovarian Carcinoma: An Overview**. *Int J Gynecol Pathol* 2008, **PAP**.

534 **Table 1**: Characteristics of the populations included in the seven analytic data sets

	TCGA	Mayo	Yoshihara *et al.*	Tothill *et al.*	Bonome *et al.*
GEO		GSE74357	GSE32062	GSE9891	GSE26712
Platform	Affy HGU1133	Agilent 4x44K	Agilent 4x44K	Affy HGU1133	Affy HGU1133
Population	United States	United States	Japan	Australia	United States
Original Sample Size	578	528	260	285	195
Analytic Sample Size[b]	499	379	256	242	185
Age [Mean (SD)]	60.0 (11.6)	62.9 (11.3)	NA	60.3 (10.3)	61.5 (11.9)
Stage					
I	10 (2%)	7 (3%)	0 (0%)	11 (5%)	0 (0%)
II	17 (4%)	11 (3%)	0 (0%)	8 (4%)	0 (0%)
III	351 (80%)	275 (73%)	202 (79%)	178 (83%)	146 (80%)
IV	63 (14%)	86 (23%)	54 (21%)	17 (8%)	36 (20%)
Grade					
2	55 (12%)	3 (1%)	130 (51%)	80 (37%)	NA
3	386 (88%)[a]	376 (99%)	126 (49%)	134 (63%)	NA
Debulking					
Optimal	325 (74%)	287 (76%)	101 (39%)	132 (62%)	89 (49%)
Suboptimal	116 (26%)	87 (23%)	155 (61%)	82 (38%)	93 (51%)

535

536 NA: Data not reported

537 [a]One sample was labeled as 'Grade 4' in TCGA

538 [b]samples without full survival data were exluded in survival analyses

Table 2: SAM moderated t score vector Pearson correlations between clusters identified using *k* = 3 versus *k* = 4 within each population.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4[a]
TCGA	0.99	0.98	0.69	0.53
Mayo	0.91	0.97	0.48	0.67
Yoshihara *et al.*	1.00	0.94	0.80	0.59
Tothill *et al.*	0.95	1.00	0.22	0.89
Bonome *et al.*	0.98	0.99	0.80	0.28

[a]Correlations for cluster 3 (*k* = 3) versus cluster 4 (*k* = 4).

Table 3: SAM moderated t score vector Pearson correlations between analogous clusters across populations[a]

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
k = 3[a]	0.77 - 0.85	0.80 - 0.90	0.65 - 0.77	NA
k = 4[a]	0.77 - 0.85	0.83 - 0.89	0.51 - 0.76	0.61 - 0.75
Bonome *k* = 3[b]	0.45 – 0.46	-0.02 - 0.12	0.22 - 0.42	NA
Bonome *k* = 4[b]	0.50 - 0.57	-0.04 - 0.04	0.13 - 0.29	0.26 - 0.43

[a]Correlation ranges for TCGA, Mayo, Yoshihara, and Tothill.

[b]Bonome is removed from gene set analyses because of low correlating clusters.

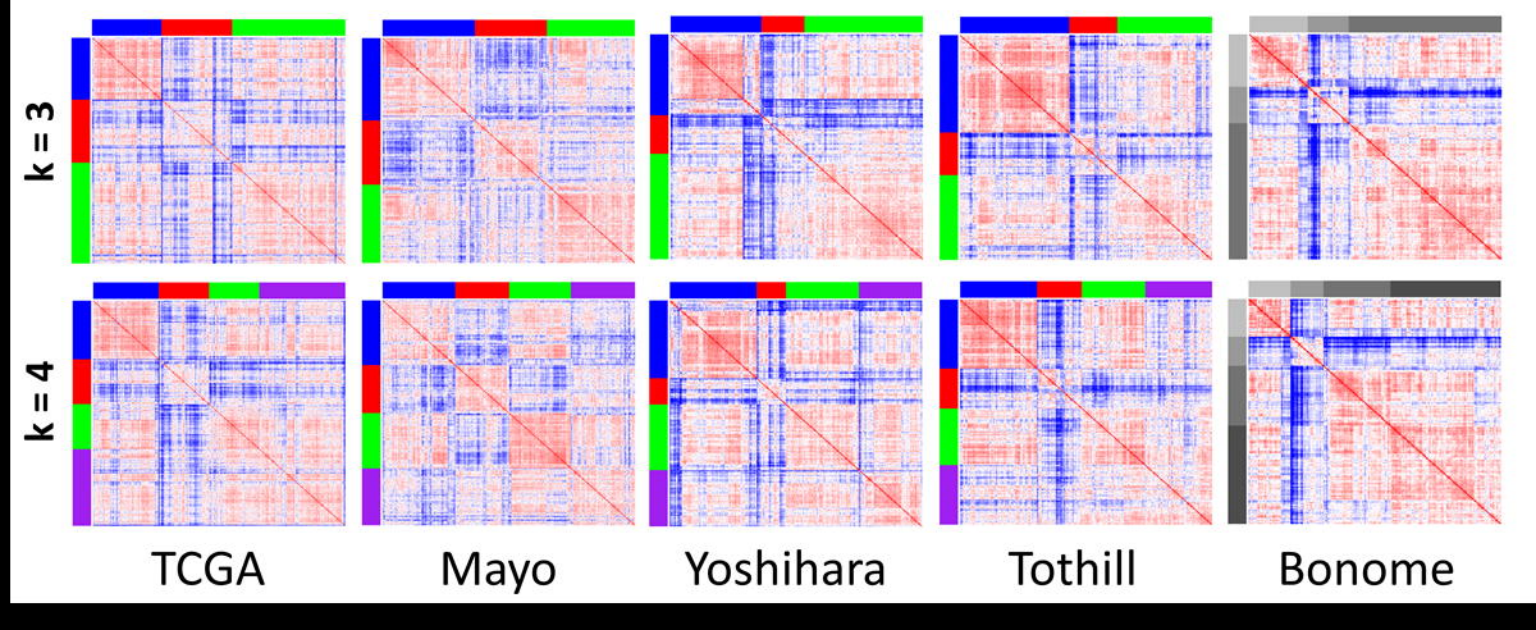
549 **Table 4:** Distributions of sample membership in the clusters identified in our study by the
550 original cluster assignments in the TCGA, Tothill, and Konecny studies. Clusters identified in
551 our study using *k*-means clustering with $k = 3$ and $k = 4$

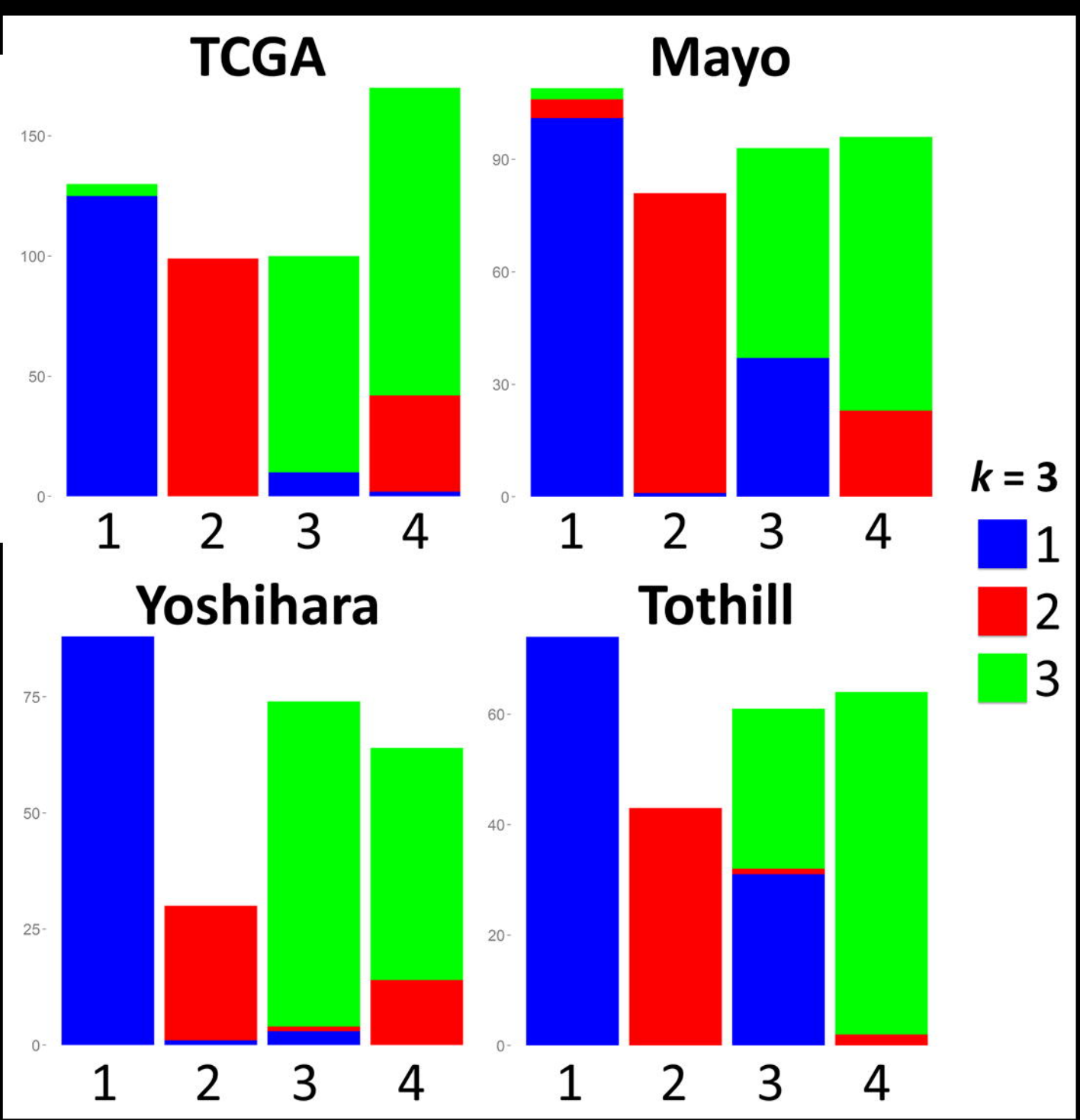
	TCGA					Tothill *et al.*							Konecny *et al.*				
	Mes	Pro	Imm	Dif	NC[a]	C1	C2	C3	C4	C5	C6	NC[a]	C1	C2	C3	C4	NA[b]
Cluster 1	98	2	20	11	6	77	22	0	0	0	0	6	16	13	2	26	82
Cluster 2	1	111	0	11	16	1	0	0	3	35	2	5	0	16	36	0	56
Cluster 3	0	21	75	106	21	0	22	6	41	0	0	22	26	31	5	0	70
	Mes	Pro	Imm	Dif	NC[a]	C1	C2	C3	C4	C5	C6	NC[a]	C1	C2	C3	C4	NA[b]
Cluster 1	97	4	12	12	5	74	0	0	0	0	0	0	7	12	3	25	62
Cluster 2	1	85	0	0	13	1	0	0	1	34	2	5	0	9	31	0	41
Cluster 3	0	5	80	3	12	3	42	0	1	1	0	14	29	6	0	1	57
Cluster 4	1	40	3	113	13	0	2	6	42	0	0	14	6	33	9	0	48

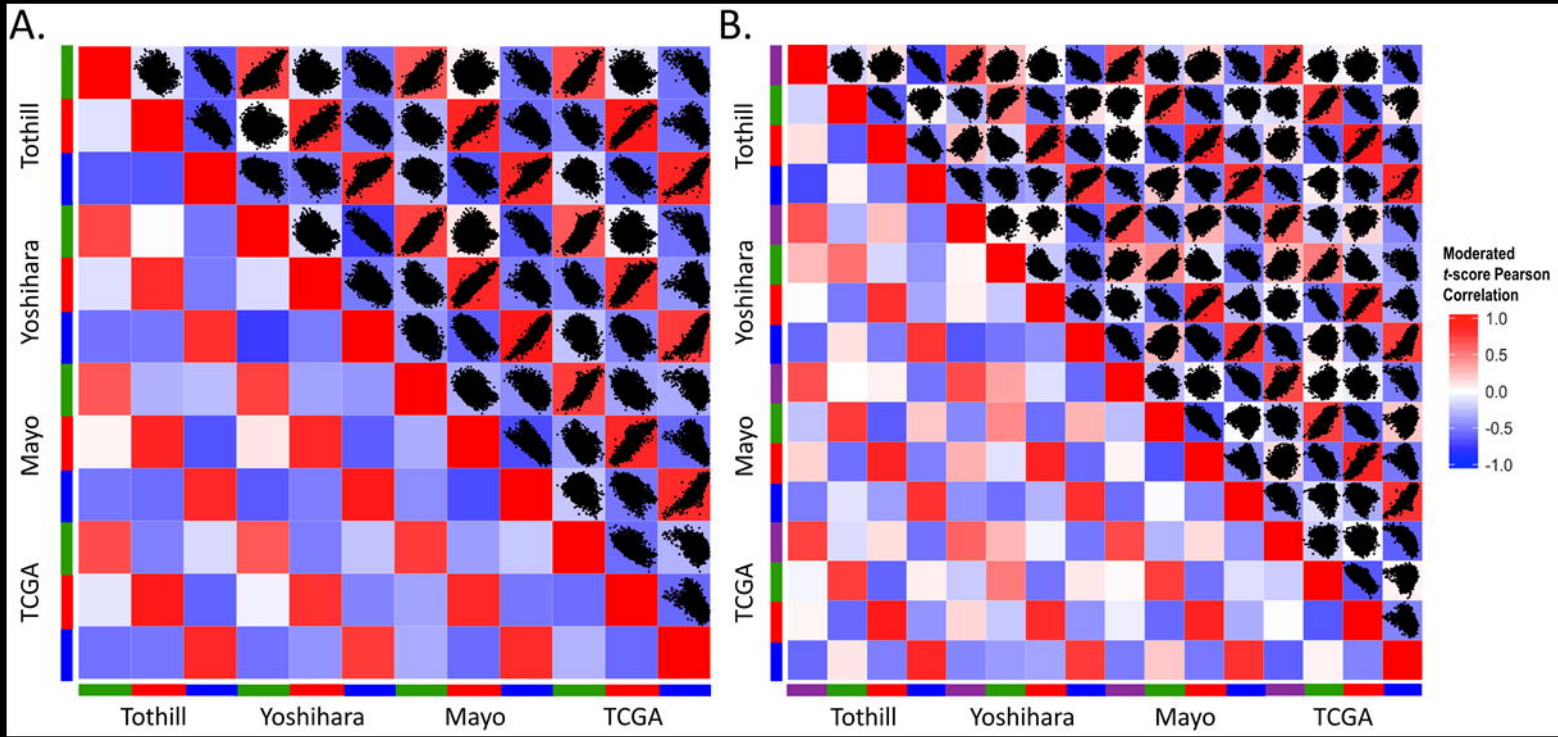
552 [a]NC = Samples not clustered in original publication

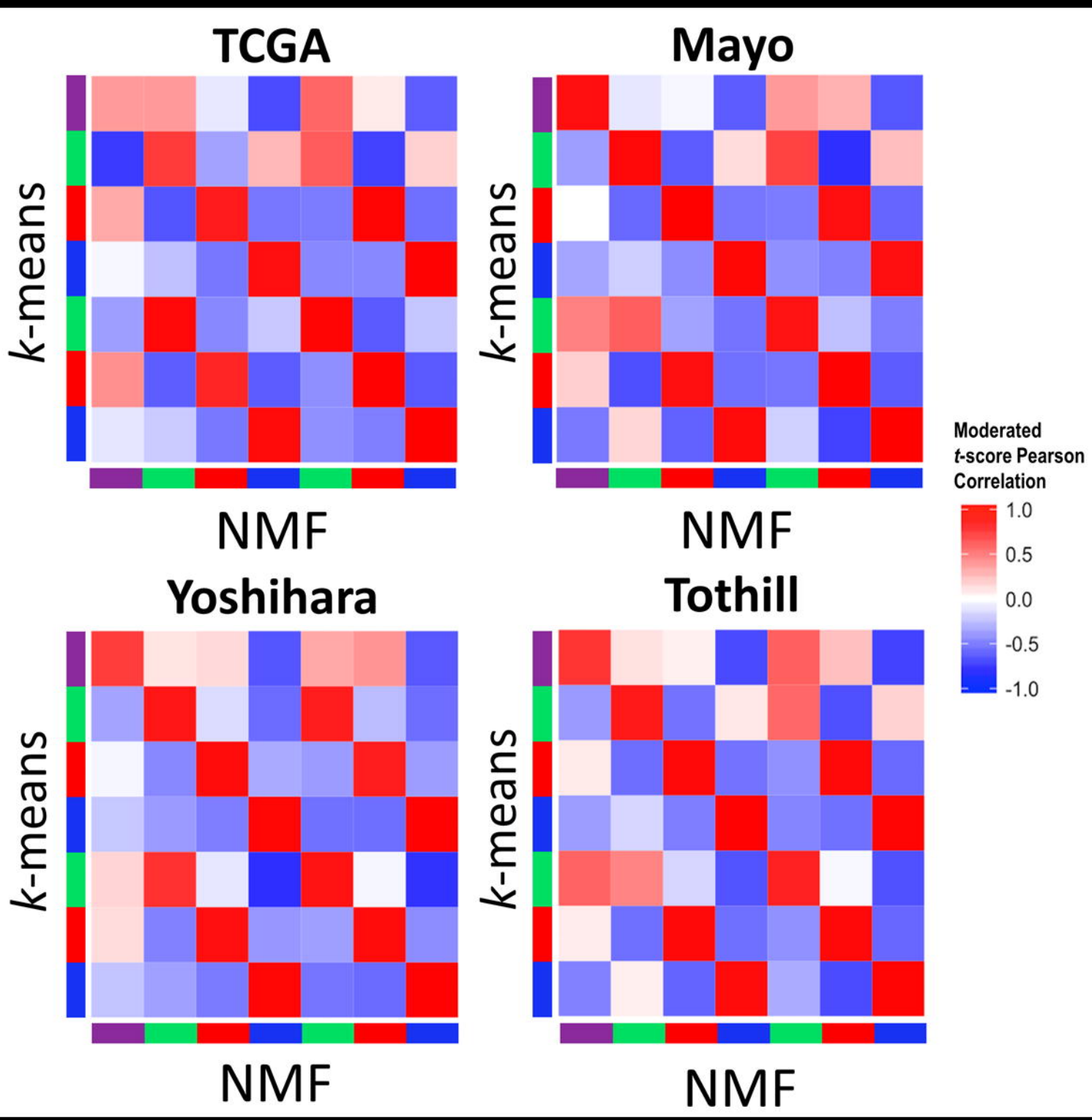
553 [b]NA = Samples not assessed at the time of the original publication

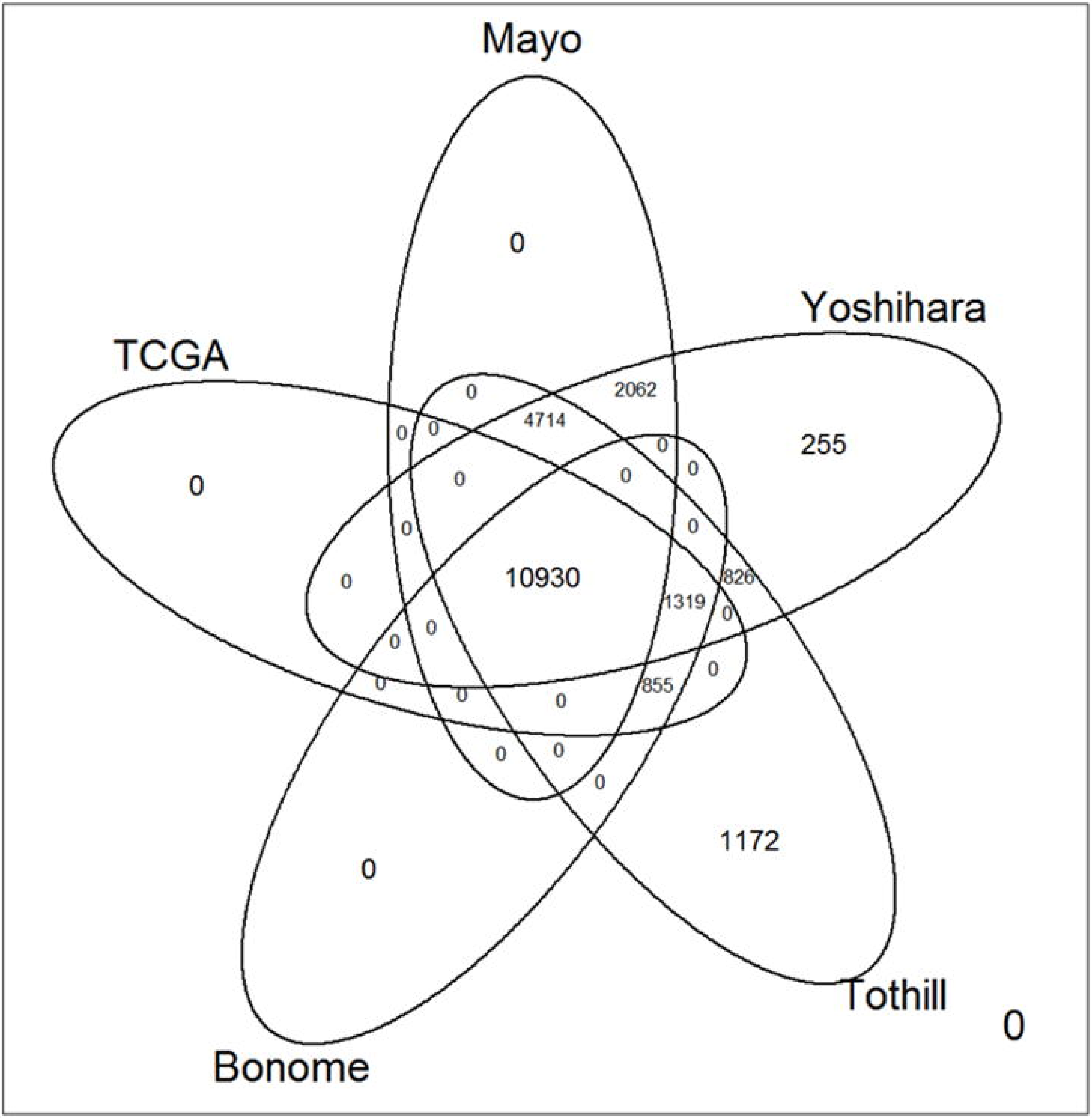
554 NOTE: The corresponding labels for the generally similar HGSC gene expression subtypes
555 observed in the TCGA, Tothill, and Konecny studies are, respectively: mesenchymal/C1/C4,
556 proliferative/C5/C3, immunoreactive/C2/C1, and differentiated/C4/C2)


k = 3
k = 4
TCGA
Mayo
Yoshihara
Tothill
Bonome


TCGA
150
100
50
0
1
2
3
4
Mayo
90
60
30
0
1
2
3
4
Yoshihara
75
50
25
0
1
2
3
4
Tothill
60
40
20
0
1
2
3
4
k = 3
1
2
3


A.
B.
Tothill
Yoshihara
Mayo
TCGA
Moderated
t-score Pearson
Correlation
1.0
0.5
0.0
-0.5
-1.0

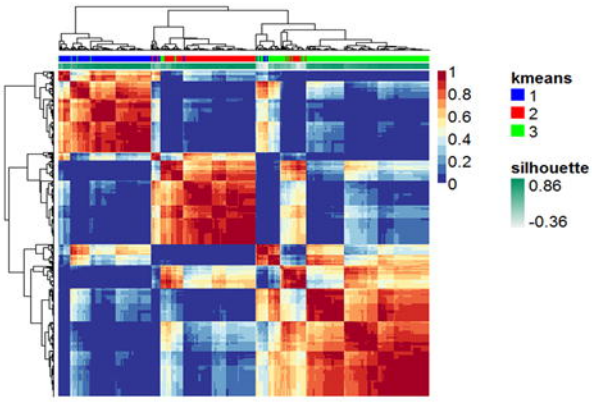

TCGA
Mayo
Yoshihara
Tothill
k-means
NMF
Moderated
t-score Pearson
Correlation
1.0
0.5
0.0
-0.5
-1.0



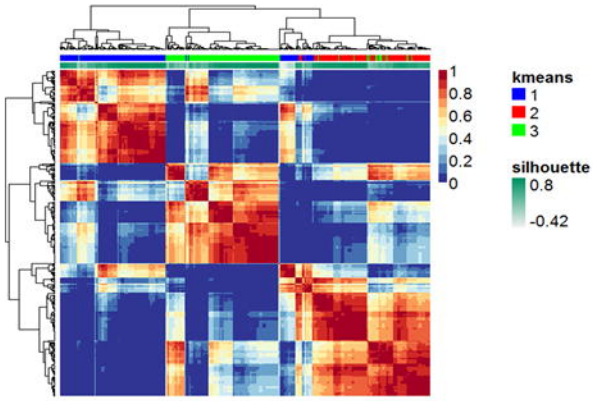
Mayo
Yoshihara
TCGA
Bonome
Tothill
0
2062
4714
255
10930
1826
1319
855
1172

A.

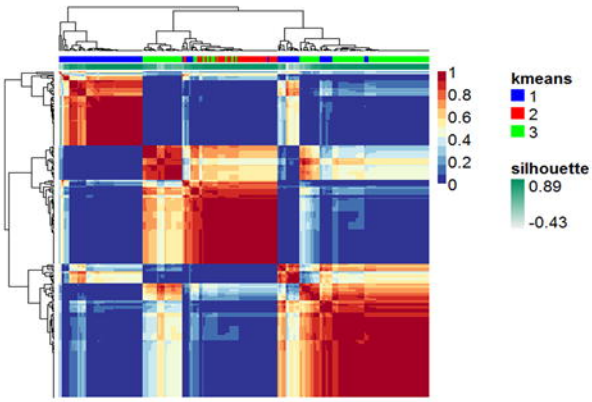
TCGA



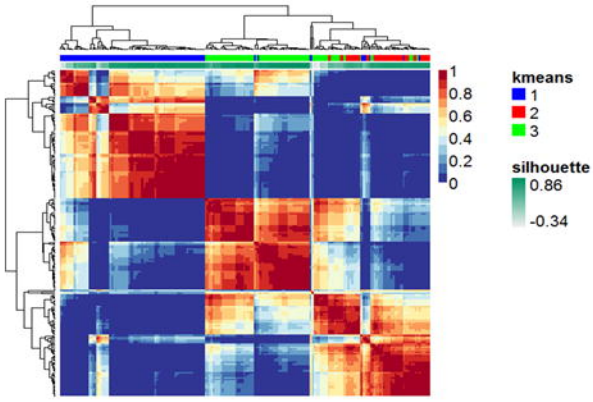
Mayo



Yoshihara

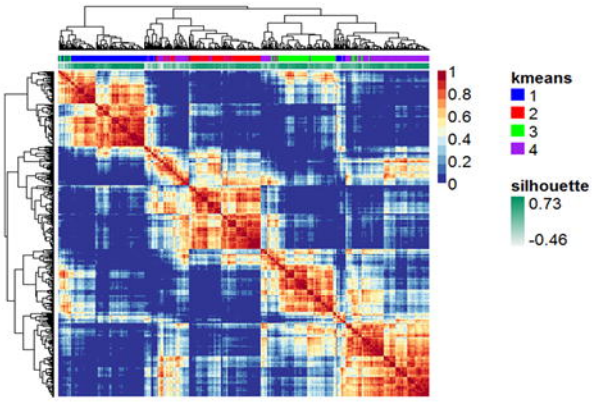


Tothill

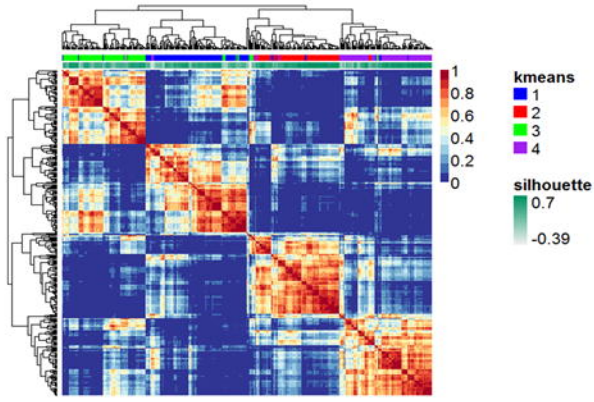


B.

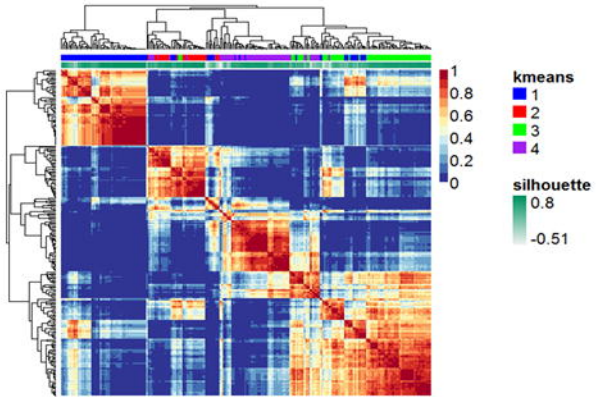
TCGA



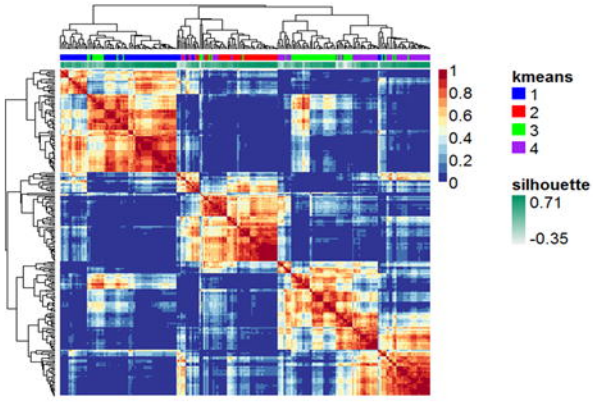
Mayo

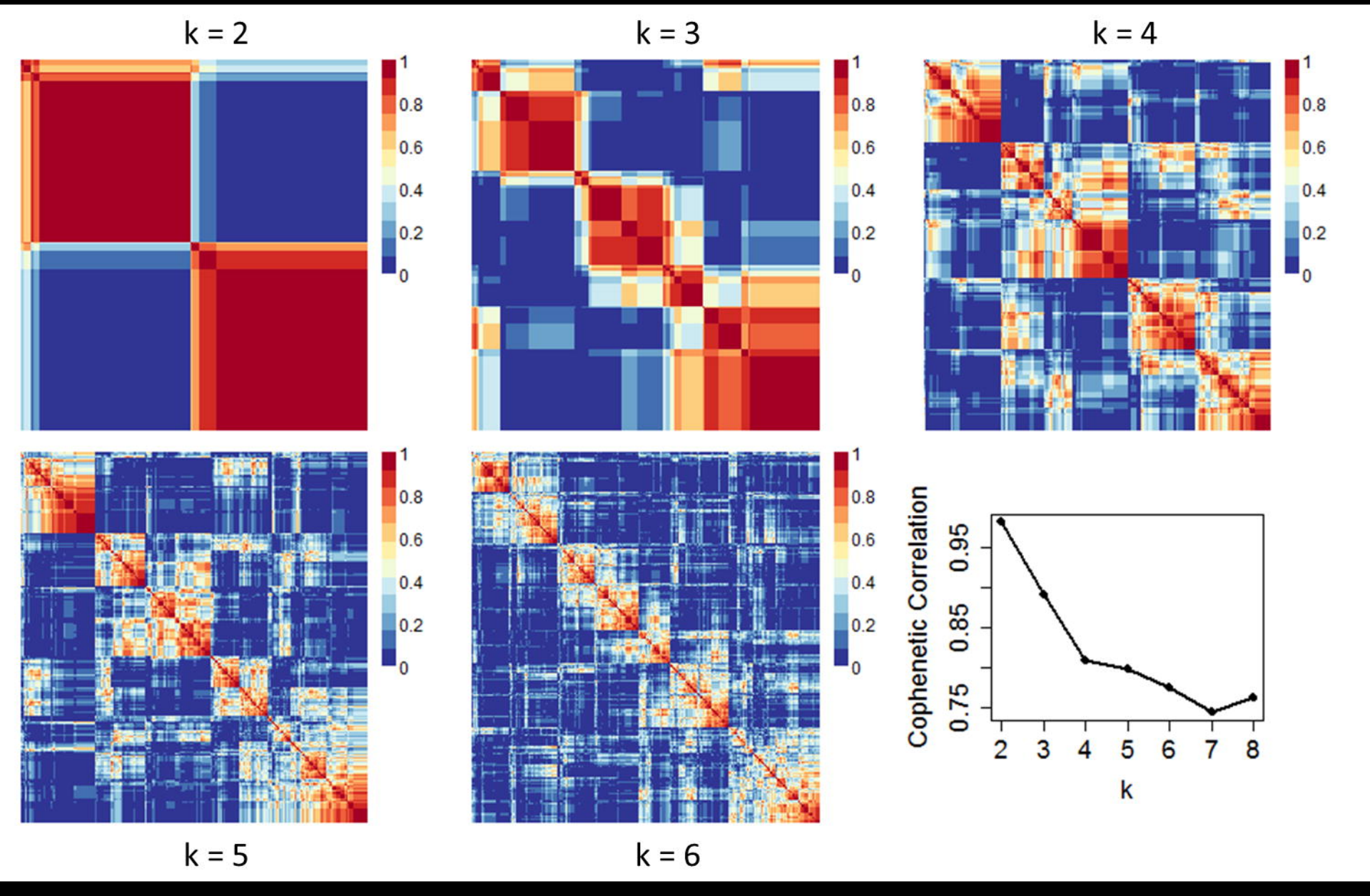


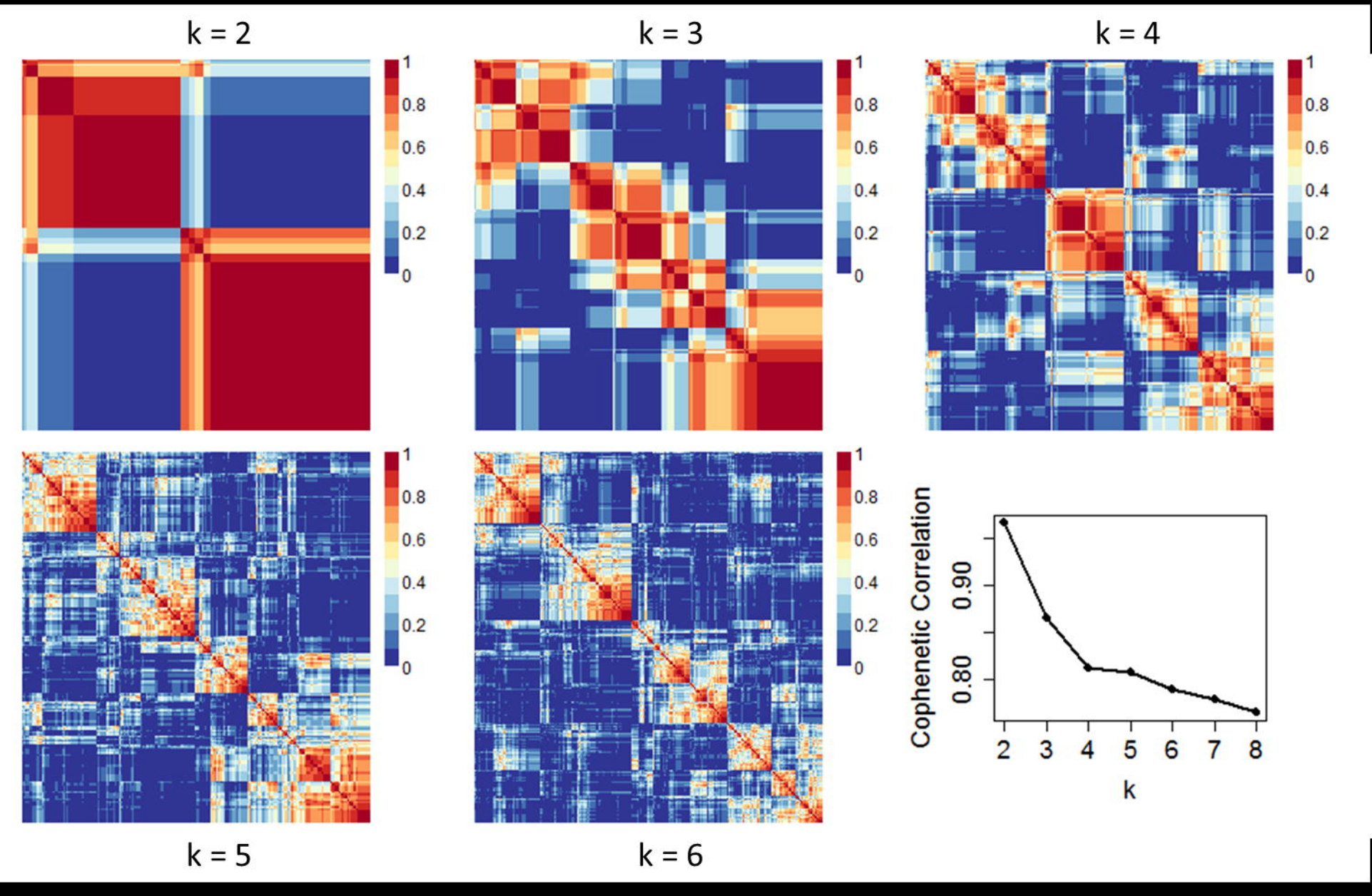
Yoshihara

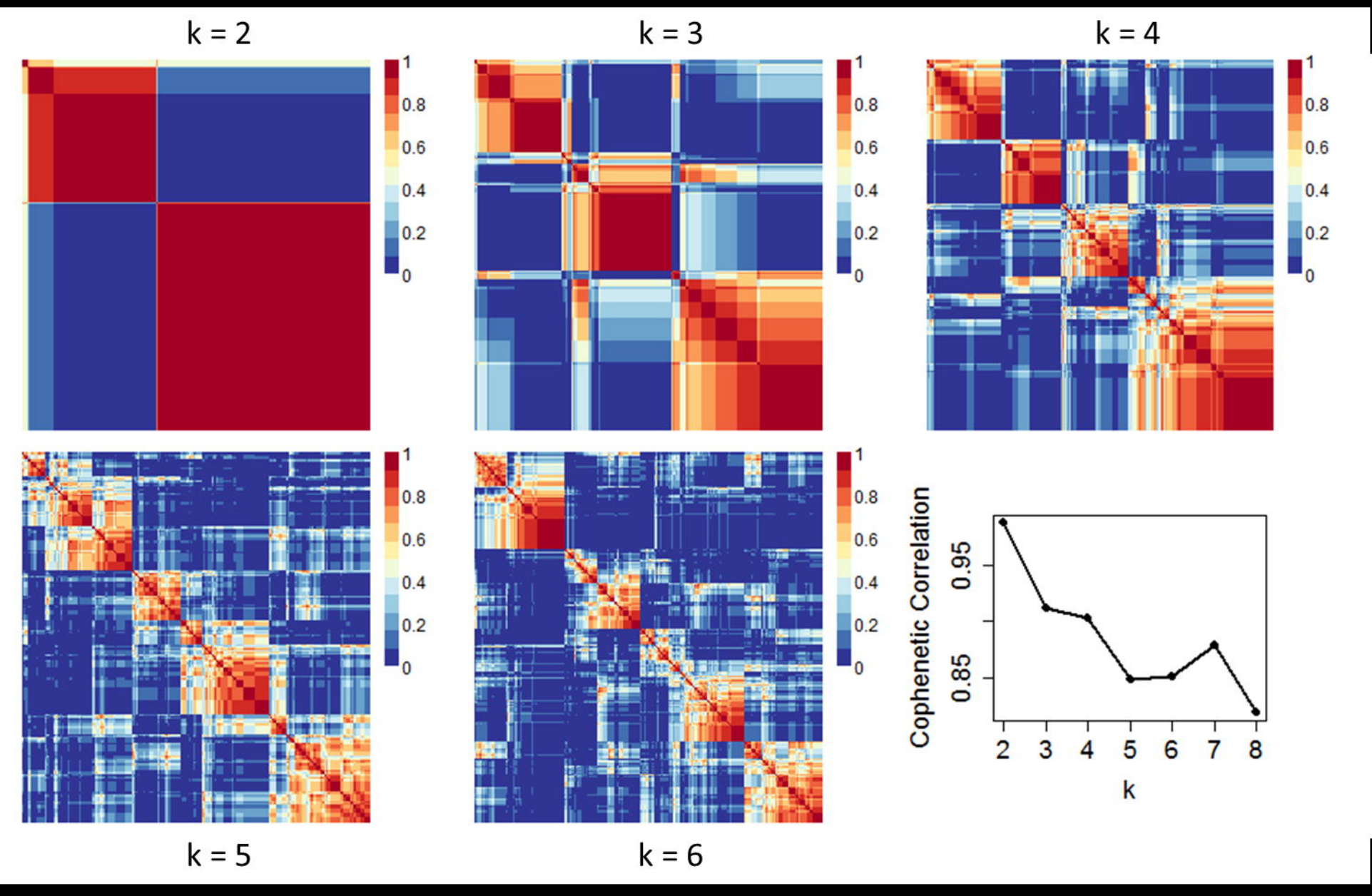


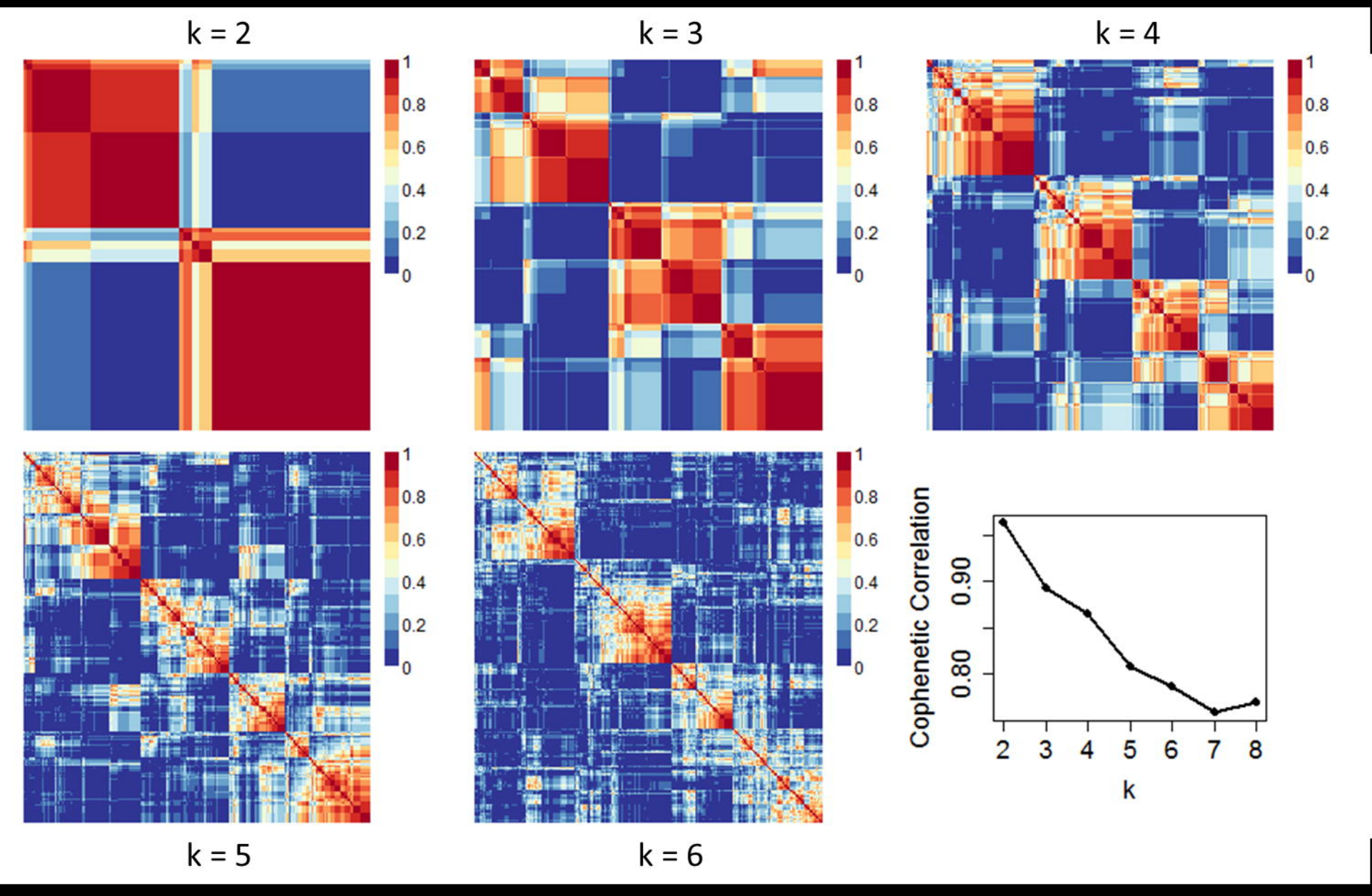
Tothill

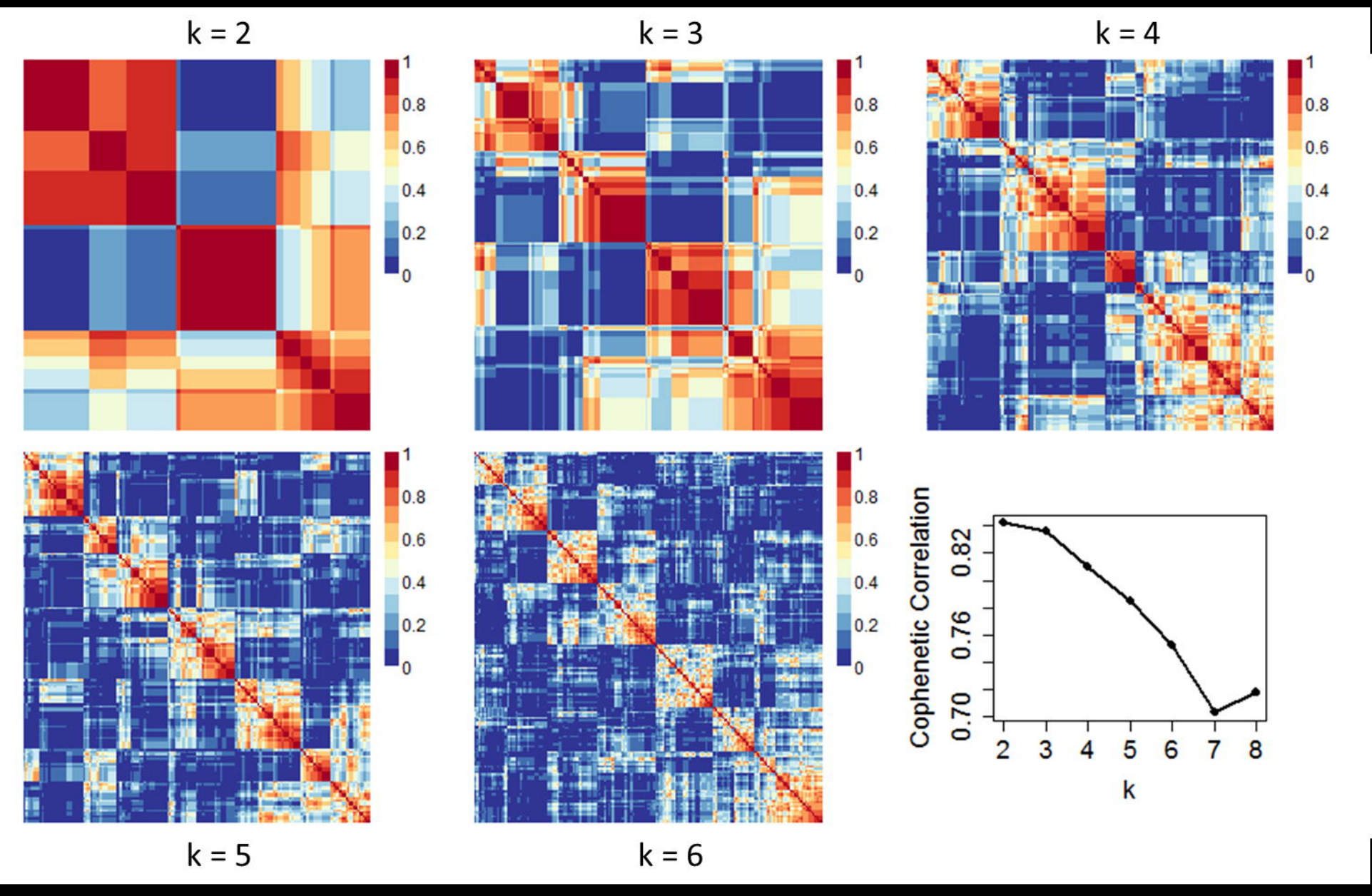


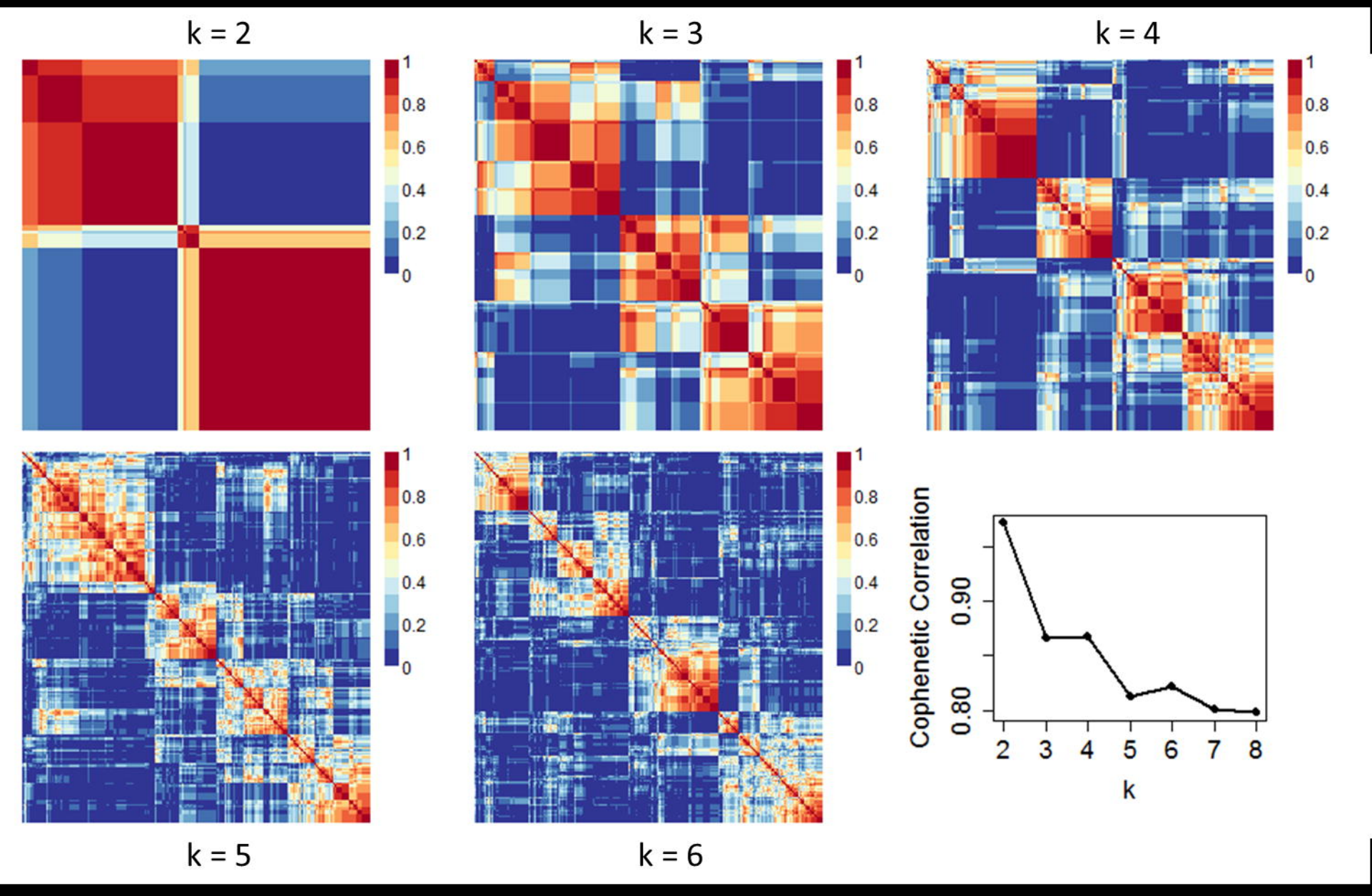

k = 2
k = 3
k = 4
k = 5
k = 6
1
0.8
0.6
0.4
0.2
0
Cophenetic Correlation
0.75
0.85
0.95
2
3
4
5
6
7
8
k

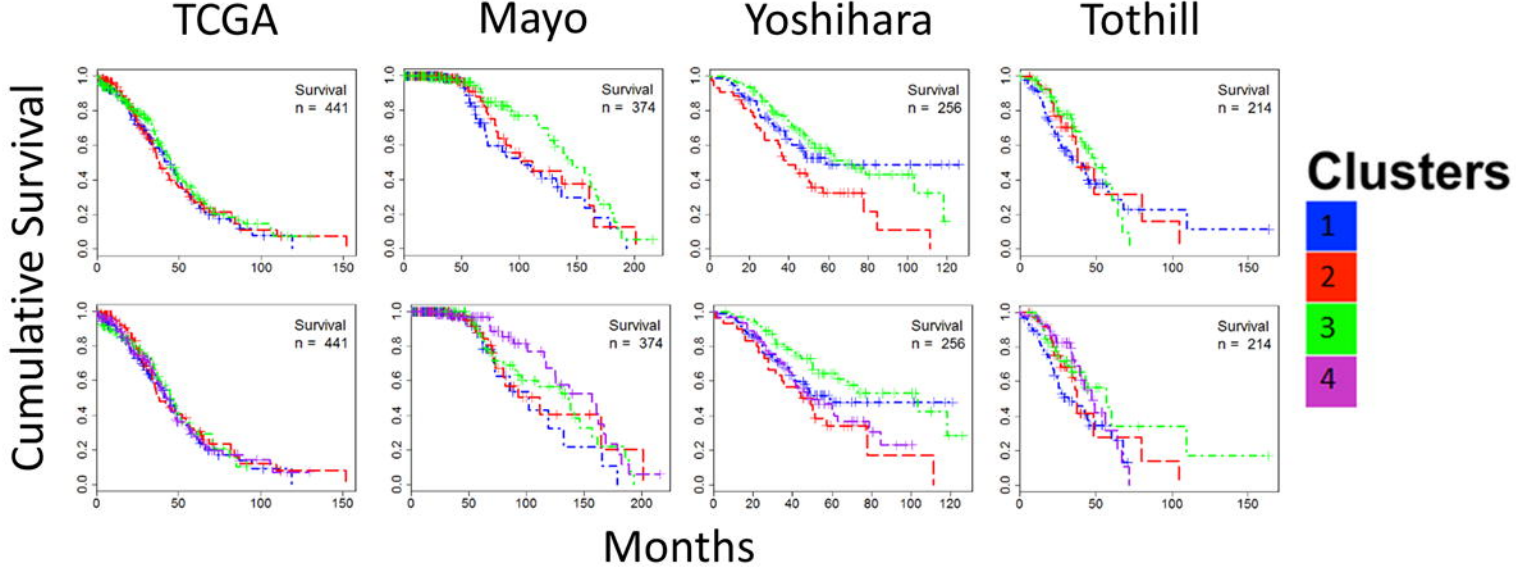

k = 2
k = 3
k = 4
k = 5
k = 6
1
0.8
0.6
0.4
0.2
0
Cophenetic Correlation
0.90
0.80
2
3
4
5
6
7
8
k


k = 2
k = 3
k = 4
k = 5
k = 6
1
0.8
0.6
0.4
0.2
0
Cophenetic Correlation
0.95
0.85
2
3
4
5
6
7
8
k


k = 2
k = 3
k = 4
k = 5
k = 6
1
0.8
0.6
0.4
0.2
0
Cophenetic Correlation
0.90
0.80
2
3
4
5
6
7
8
k


k = 2
k = 3
k = 4
k = 5
k = 6
1
0.8
0.6
0.4
0.2
0
Cophenetic Correlation
0.70
0.76
0.82
2
3
4
5
6
7
8
k


k = 2
k = 3
k = 4
k = 5
k = 6
1
0.8
0.6
0.4
0.2
0
Cophenetic Correlation
0.90
0.80
2
3
4
5
6
7
8
k


TCGA
Mayo
Yoshihara
Tothill
Cumulative Survival
Survival
n = 441
n = 374
n = 256
n = 214
Months
Clusters
1
2
3
4