1   TITLE:

2   Comprehensive cross-population analysis of high-grade serous ovarian cancer supports

3   no more than three subtypes

4   AUTHORS:

5   Gregory P. Way[*,†,‡], James Rudd[‡,§], Chen Wang[**,·], Habib Hamidi[††], Brooke L. Fridley[‡‡],

6   Gottfried Konecny[††], Ellen L. Goode[**], Casey S. Greene[†,‡,§§,1], Jennifer A. Doherty[§,**,2]

7   AFFILIATIONS:

8   [*]Genomics and Computational Biology Graduate Program, University of Pennsylvania,

9   Philadelphia, PA 19103, USA

10   [†]Department of Systems Pharmacology and Translational Therapeutics, Perelman School

11   of Medicine, University of Pennsylvania, Philadelphia, PA 19103, USA

12   [‡]Quantitative Biomedical Sciences, Norris Cotton Cancer Center, Geisel School of

13   Medicine at Dartmouth, Lebanon, NH 03766, USA

14   [§]Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH

15   03766, USA

16   [**]Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA

17   [††]Department of Medicine, David Geffen School of Medicine, University of California,

18   Los Angeles, CA 90095, USA

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

20   66160, USA

21   [§§§]Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH

22   03755, USA

23

24 <u>CO-CORRESPONDING AUTHORS</u>:

25 [1]10-131 SCTR 34th and Civic Center Blvd, Philadelphia, PA 19104; Phone: 215-573-

26 2991; Fax: 215-573-9135; csgreene@upenn.edu

27 [2]One Medical Center Drive, Lebanon, NH 03766; Phone: 603-653-9065; Fax: 603-653-

28 9093; Jennifer.A.Doherty@Dartmouth.edu

29

30 <u>REFERENCE NUMBERS</u>:

31   Gene Expression Omnibus Data Accession: GSE74357

32

33 <u>CONFLICTS OF INTEREST</u>:

34   The authors do not declare any conflicts of interest.

35

36 <u>OTHER PRESENTATIONS</u>:

37   Aspects of this study were presented at the 2015 AACR Conference and the 2015 Rocky

38   Mountain Bioinformatics Conference.

39

40 <u>RUNNING TITLE (35 characters with spaces)</u>:

41   *Three or less HG serous OV subtypes*

42

43 <u>KEYWORDS</u>:

44   Ovarian Cancer; Molecular Subtypes; Unsupervised Clustering; Reproducibility

45

46

ABSTRACT:

47

48   Four gene expression subtypes of high-grade serous ovarian cancer (HGSC) have been

49   previously described. In these studies, a fraction of samples that did not fit well into the four

50   subtype classifications were excluded. Therefore, we sought to systematically determine the

51   concordance of transcriptomic HGSC subtypes across populations without removing any

52   samples. We created a bioinformatics pipeline to independently cluster the five largest mRNA

53   expression datasets using k-means and non-negative matrix factorization (NMF). We

54   summarized differential expression patterns to compare clusters across studies. While previous

55   studies reported four subtypes, our cross-population comparison does not support four. Because

56   these results contrast with previous reports, we attempted to reproduce analyses performed in

57   those studies. Our results suggest that early results favoring four subtypes may have been driven

58   by including serous borderline tumors. In summary, our analysis suggests that either two or

59   three, but not four, gene expression subtypes are most consistent across datasets.

60

61

62   INTRODUCTION:

63   Invasive ovarian cancer is a heterogeneous disease typically diagnosed at a late stage,

64   with high mortality (Kurman and Shih 2010) . The most aggressive and common histologic type

65   is high-grade serous (HGSC) (Vang *et al.* 2009), characterized by extensive copy number

66   variation and *TP53* mutation (The Cancer Genome Atlas 2011). Given the genomic complexity

67   of these tumors, mRNA expression can be thought of as a summary measurement of these

68   genomic and epigenetic alterations, to the extent that the alterations influence gene expression in

69   either the cancer or stroma.

3

70   Four gene expression subtypes with varying components of mesenchymal, proliferative,

71   immunoreactive, and differentiated gene expression signatures have been reported in all studies

72   of HGSC to date (Bonome *et al.* 2008; Tothill *et al.* 2008; The Cancer Genome Atlas 2011; Tan

73   *et al.* 2013; Konecny *et al.* 2014). Two of these also observed survival differences across

74   subtypes (Tothill *et al.* 2008; Konecny *et al.* 2014). Tothill *et al.* first identified four HGSC

75   subtypes (as well as two other subtypes which largely included low grade serous and serous

76   borderline tumors) in an Australian population using *k*-means clustering. Later, The Cancer

77   Genome Atlas (TCGA) used non-negative matrix factorization (NMF) and also reported four

78   subtypes which were labeled as: 'mesenchymal', 'differentiated', 'proliferative', and

79   'immunoreactive' (The Cancer Genome Atlas 2011). The TCGA group also applied NMF

80   clustering to the Tothill data, and observed concordance with four subtypes (The Cancer Genome

81   Atlas 2011). Konecny *et al.* applied NMF to cluster an independent set of HGSC samples and

82   reported four subtypes, which they labeled as C1-C4 (Konecny *et al.* 2014). These subtypes were

83   similar to those in the TCGA but a subtype classifier trained on these subtypes better

84   differentiated survival in their own data, and in data from TCGA and Bonome *et al.* (Bonome *et*

85   *al.* 2008).

86   Despite this extensive research in the area, work to date has several limitations. In both

87   TCGA and Tothill *et al.*, ~8-15% of samples were excluded from analyses. A reanalysis of the

88   TCGA data showed that over 80% of the samples could be assigned to more than one subtype

89   (Verhaak *et al.* 2012). In more recent TCGA analyses by the Broad Institute Genome Data

90   Analysis Center (GDAC) Firehose initiative with the largest number of HGSC cases evaluated to

91   date (n = 569), three subtypes fit the data better than four (Broad Institute TCGA Genome Data

4

92   Analysis Center 2014, 2015). This uncertainty in HGSC subtyping led us to determine if four

93   homogeneous subtypes exist across study populations.

94   To comprehensively characterize subtypes, we analyze data from the five largest

95   independent studies to date, including our own collection of samples, using a standardized

96   bioinformatics pipeline. We apply *k*-means clustering as well as NMF to each population without

97   removing "hard-to-classify" samples. Our goal is to rigorously assess the number of subtypes.

98   These independent and parallel within-dataset analyses followed by cross-dataset comparison

99   sidestep gene expression platform or dataset biases that could affect clustering if under or

100   overcorrected. This contrasts with earlier work that pooled datasets together to identify subtypes

101   (Tan *et al.* 2013) and ensures that subtypes identified are not induced by dataset or batch effects.

102   We summarize each subtype's expression patterns and comprehensively characterize correlations

103   between subtype-specific gene expression across populations.

104   Our cross-population comparative analysis does not support that four HGSC subtypes

105   exist; rather the data more strongly support an interpretation that there are either two or three

106   subtypes. We show that the support for four subtypes observed in TCGA's reanalysis of Tothill

107   *et al.* (The Cancer Genome Atlas 2011) is lost when serous borderline tumors, which have very

108   different genomic profiles and survival than HGSC (Ouellet *et al.* 2005; Bonome *et al.* 2005), are

109   excluded before clustering. Our work also highlights the impact that a single study can have on

110   the trajectory of subtyping research and suggests the importance of periodic histopathologic

111   review and rigorous reanalysis of existing data for cross-study commonalities.

112

113   <u>MATERIALS AND METHODS:</u>

114   *Data inclusion*

5

115   We applied inclusion criteria as described in the supplementary materials using data from

116   the R package, curatedOvarianData (Ganzfried *et al.* 2013) and our own novel dataset ("Mayo")

117   (Konecny *et al.* 2014) (Supplementary Table S1). These criteria selected HGCS samples that

118   were not duplicates from studies including at least 130 HGSC cases assayed on standard

119   microarrays. Data from the new Mayo HGSC samples as well as other samples with mixed

120   histologies and grades, for a total of 528 additional ovarian tumor samples, was deposited in

121   NCBI's Gene Expression Omnibus (GEO) (Edgar *et al.* 2002); these data can be accessed with

122   the accession number GSE74357

123   (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74357). All study participants

124   provided written informed consent, and this work was approved by the Mayo Clinic and

125   Dartmouth College Institutional Review Boards.

126   After applying the unified inclusion criteria, our final analytic datasets include: TCGA (n

127   = 499) (The Cancer Genome Atlas 2011; Broad Institute TCGA Genome Data Analysis Center

128   2014); Mayo (n = 379; GSE74357) (Konecny *et al.* 2014); Yoshihara (n = 256;

129   GSE32062.GPL6480) (Yoshihara *et al.* 2012); Tothill (n = 242; GSE9891) (Tothill *et al.* 2008);

130   and Bonome (n = 185; GSE26712) (Bonome *et al.* 2008) (Table 1). We restricted analyses to the

131   10,930 genes measured successfully in all five populations (Supplementary Fig. S1).

132

133   *Clustering*

134   We performed independent clustering within each dataset to avoid potential biases from

135   different platforms or studies. As detailed in the Supplementary Methods, we identified the

136   1,500 genes with the highest variance from each dataset and used the union of these genes (n =

137   3,698) for clustering. We performed clustering within each dataset using each potential *k* from 2-

138 8 clusters. We performed k-means clustering in each population using the R package "cluster"

139 (version 2.0.1) (Maechler *et al.* 2014) with 20 initializations. We repeated these analyses using

140 NMF in the R package "NMF" (version 0.20.5) (Brunet *et al.* 2004) with 100 different random

141 initializations for each k. As done in prior studies, we calculated cophenetic correlation

142 coefficients to select appropriate k for each dataset after NMF clustering with 10 consensus runs

143 for $k = 2$ through 8.

144

145 *Identification of analogous clusters within and across studies*

146 We performed significance analysis of microarray (SAM) (Tusher *et al.* 2001; Schwender

147 *et al.* 2006) analysis on all clusters from each study using all 10,930 genes. This resulted in a

148 cluster-specific moderated t statistic for each of the input genes (Schwender 2012). To

149 summarize the expression patterns of all 10,930 genes for a specific cluster in a specific

150 population, we combined gene-wise moderated t statistics into a vector of length 10,930. The

151 TCGA subtype labels have become widely used in the field. To generate comparable labels

152 across k and across studies, we mapped our TCGA subtype assignments back to the original

153 TCGA labels to define reference clusters at $k = 4$ (that is, mesenchymal-like, proliferative-like,

154 etc.). Clusters in other populations that were most strongly correlated with the TCGA clusters

155 were assigned the same label.

156

157 *Clustering analysis of randomized data*

158 Any clustering procedure is expected to induce strong correlational structure across

159 clusters within a dataset even if there is no true underlying structure. However, if there is no true

160 underlying structure, clusters across datasets are not expected to be correlated. To assess this, we

7

161   used the same datasets but shuffled each gene's expression vector to disrupt the correlative

162   structure. We performed within and cross-study analyses of cluster identification using this set of

163   data that were parallel to those performed using the non-randomized data.

164

165   *Assessing the reproducibility of single-population studies*

166   We compared our sample assignments at $k = 2 - 4$ to the four subtypes reported in the

167   Tothill, TCGA, and Konecny publications (Tothill *et al.* 2008; The Cancer Genome Atlas 2011;

168   Konecny *et al.* 2014). Because the labels that were assigned in TCGA's reanalysis of the Tothill

169   data were not available, we performed NMF consensus clustering of Tothill's data without

170   removing LMP samples in order to generate labels for comparison.

171

172   *Data Availability*

173   We provide software to download the required data and reproduce our analyses. The

174   software is provided under a permissive open source license (Way *et al.* 2015). Analyses were

175   run in a Docker container, allowing the computing environment to be recreated (Boettiger 2015).

176   Our Docker image can be pulled from here: https://hub.docker.com/r/gregway/hgsc_subtypes/.

177   This allows interested users to freely download the software, reproduce the analyses, and then

178   build on this work. All data used in this analysis is publicly available including data we

179   generated (accessible under GEO accession GSE74357).

180

181   RESULTS:

182   *Clustering*

183   To visually inspect the consistency and distinctness of clusters, we compared sample-by-

184   sample correlation heatmaps. For $k = 2$ to 4 within each study, we observed high sample-by-

185   sample correlations within clusters and relatively low sample-by-sample correlations across

186   clusters (Supplementary Fig. S2). Clustering results using NMF were similar to k means results

187   (Supplementary Fig S3).

188

189   *Correlation of cluster-specific expression patterns*

190   Across datasets, we observed strong positive correlations of moderated t score vectors

191   between analogous clusters in TCGA, Tothill, Mayo, and Yoshihara (Fig. 1; Table 2). However,

192   clustering of the Bonome data did not correlate strongly with clusters identified in the other

193   datasets (Table 2). We believe that we were unable to assign parallel subtypes in Bonome

194   because of either RNA contamination or inappropriate grading assignments. However, more

195   work is required in order to identify exactly why we were unable to classify.

196   To assess our analytical approach, we performed an analysis using randomized data. This

197   showed that within-population correlation structure was induced by clustering, but structure

198   between populations was not (Supplementary Fig. S4). Comparing Figure 1 with S4, we

199   observed much higher correlation across datasets (Fig. 1), which was lost after randomization

200   (Supplementary Fig. S4). For example, for $k = 2$, the TCGA and Mayo cluster correlations for

201   analogous clusters was high (top left panel in Fig. 1). Conversely, the same relationship in

202   randomized data (second row, first column panel in Supplementary Fig. S4) showed correlations

203   near zero. This indicates that the high correlations observed across datasets in Figure 1 are

204   induced by similar underlying structure in the data.

205   Across studies, positive correlations between analogous clusters and negative correlations

206 between non-analogous clusters were stronger for clusters identified when $k = 2$ and $k = 3$ than

207 when $k = 4$ (Fig. 1), with comparable statistical precision (Supplementary Table S2). These

208 cross-population comparisons suggested that two and three subtypes fit HGSC gene expression

209 data more consistently than the four widely accepted subtypes.

210   Within each population, clusters identified by NMF were similar to those identified using

211 k-means clustering (Fig. 2) suggesting that these results were independent of clustering

212 algorithm. With NMF, both positive and negative correlations were stronger for $k = 2$ and $k = 3$

213 than for $k = 4$. Across $k = 3$ and $k = 4$, correlations were strongest for clusters 1 and 2. Sample

214 cluster assignments for both k-means and NMF clusters are provided in Supplementary Table S3.

215

216 *Comparison with previously-identified HGSC clusters*

217   Our clustering results for the Tothill, TCGA, and Mayo datasets were highly concordant

218 with the clustering described in the original publications (Tothill *et al.* 2008; The Cancer

219 Genome Atlas 2011; Konecny *et al.* 2014), as evidenced by the high degree of consistent overlap

220 in sample assignments to the previously-defined clusters (Table 3). Our cross-study cluster 1 was

221 mostly mapped to the "Mesenchymal" label from TCGA, "C1" from Tothill, and "C4" from

222 Mayo. This cluster was the most stable in our analysis within all datasets, across $k = 2$, 3 and 4,

223 and across clustering algorithms. Cross-study cluster 2, which was also observed consistently,

224 was most similar to the "Proliferative" label from TCGA, "C5" from Tothill, and "C3" from

225 Mayo. Cross-study cluster 3 for $k = 3$ was associated with both the "Immunoreactive" and

226 "Differentiated" TCGA labels, "C2" and "C4" in Tothill, and "C1" and "C2" in Mayo. For

227 analyses where $k = 4$, the third cluster was associated with "Immunoreactive", "C2", and "C1"

10

228   while the fourth cluster was associated with "Differentiated", "C4", and "C2" for TCGA, Tothill,

229   and Mayo respectively.

230

231   *Meta-research into previous HGSC subtyping studies*

232   Each of the publications that only considered high-grade samples (TCGA and Konecny *et*

233   *al.*) found clustering coefficients consistent with $k = 2$, $k = 3$, and $k = 4$. Nevertheless, each

234   publication concludes the existence of four subtypes, while our cross-population analysis

235   suggested that two or three clusters fit HGSC data better than four clusters.

236   To compare with previous results, we evaluated the number of subtypes that fit the data

237   best within each study by calculating cophenetic correlation coefficients at $k = 2$ through k=8

238   clusters inclusively. We observed a similar pattern in each population (Supplementary Fig S5 –

239   S7; Fig. 3A) in which the highest cophenetic correlation was reached for two clusters and, based

240   on the heatmaps, appeared to have the highest consensus. In every dataset, four clusters were not

241   observed to represent the data better than two or three. The only results in previous studies that

242   contradicted this work were from TCGA's reanalysis of the Tothill data. According to

243   supplemental figure S6.2 in the TCGA paper, the reanalysis included serous borderline tumors

244   (i.e., tumors with low malignant potential) (n = 18). The inclusion of these tumors in the TCGA

245   HGSC analyses was done even though, in the original Tothill paper, the serous borderline tumors

246   had a unique gene expression patterns and clustered entirely in a group labeled "C3".

247   To assess the extent to which serous borderline tumors inclusion drove the TCGA results,

248   we reproduced TCGA's reanalysis of Tothill *et al.*, including the serous borderline tumors (n =

249   18); we indeed observed that the cophenetic correlation is higher for $k = 4$ than $k = 3$ (Fig. 3A).

250   However, when we appropriately removed these serous borderline tumors we observed an

11

251   increase in the $k = 3$ cophenetic correlation (Fig. 3B). The results that support four subtypes were

252   generated during clustering of HGSC and serous borderline tumors combined. Subtyping

253   analyses of HGSC alone reveal less than four subtypes. Even after subtyping there remains a

254   complex and nuanced portrait of the disease.

255

256   DISCUSSION:

257   Although prior studies have reported the existence of four molecular subtypes of HGSC

258   ovarian cancer (The Cancer Genome Atlas 2011; Tothill *et al.* 2008; Konecny *et al.* 2014; Broad

259   Institute TCGA Genome Data Analysis Center 2014), our analysis suggests the existence of only

260   two or three subtypes. This conclusion is based on our observation that concordance of

261   analogous subtypes across study populations was stronger for two or three clusters as opposed to

262   four. Previous studies used either k-means or NMF clustering, and because our results

263   contradicted prior work, we performed analyses using both of these methods. Results for each

264   population were similar for the k means and NMF clustering algorithms suggesting that the

265   clustering algorithm did not drive the observed differences.

266   Because cross-population comparisons suggest that two and three clusters show more

267   consistency than four, we explored within-study heuristics (cophenetic correlation coefficients)

268   that suggested four subtypes in previous research. The cophenetic coefficient measures how

269   precisely a dendrogram retains sample by sample pairwise distances and can be used to compare

270   clustering accuracy (Sokal and Rohlf 1962). While both Konecny and TCGA reported four

271   subtypes, in both analyses $k = 2$ and $k = 3$ resulted in higher cophenetic coefficients than $k = 4$

272   (Konecny Figure 2A and TCGA Figure S6.1) (The Cancer Genome Atlas 2011), (Konecny *et al.*

273   2014). We observed the same patterns in our own reanalysis of TCGA and analysis of the

12

274   expanded Mayo cohort (Supplementary Figs. S5 and S6). Yoshihara and Tothill did not report

275   cophenetic coefficients, but our analysis of each (Supplementary Fig S7 and Fig 3A) revealed

276   similar patterns to TCGA and Konecny.

277   In the previous literature, the only report to suggest that three subtypes were

278   inappropriate was TCGA's reanalysis of the Tothill *et al.* data (supplemental Figure S6.2 in their

279   publication); the cophenetic coefficient dropped dramatically at $k = 3$ before recovering at $k = 4$

280   (The Cancer Genome Atlas 2011). Notably, TCGA's figure legend for this supplemental result

281   indicates that they did not remove serous borderline tumors from the Tothill data. Our analysis of

282   Tothill *et al.* differed from TCGA's in that we excluded serous borderline tumors and instead

283   supports the existence of two or three subtypes. To evaluate the influence of these serous

284   borderline tumors in the Tothill data, we repeated our analyses including serous borderline

285   tumors, and observed a drop in the cophenetic coefficient for $k = 3$ relative to $k = 4$ (Fig. 3). This

286   suggests that the four subtypes observed in TCGA's analysis of the Tothill data may be due, in

287   part, to the inclusion of serous borderline tumors.

288   There are several limitations to note in the HGSC data we analyzed. Given the intra-

289   tumor heterogeneity that is likely to exist (Blagden 2015), our approach would be strengthened

290   by having data on multiple areas of the tumors. Additionally, since histology and grade

291   classification have changed over time (Silverberg 2000; Soslow 2008), it is unclear whether the

292   populations we studied used comparable guidelines to determine histology and grade. We

293   attempted to exclude all low grade serous and low grade endometrioid samples because they

294   often have very different gene expression patterns and more favorable survival compared to their

295   higher grade counterparts (Vang *et al.* 2009). While the Bonome publication specified that they

296   included only high-grade tumors, grade is not included in the Bonome GSE26712 data set, so we

13

297 were unable to determine whether the grade distribution differs from the other studies (Bonome

298 *et al.* 2008). It is unclear why the Bonome clusters did not correspond to the clusters observed in

299 other populations. Lack of consistency could result from a different distribution of grade or other

300 unreported biological differences.

301 In summary, our study demonstrates that two clusters of HGSC, "mesenchymal-like" and

302 "proliferative-like", are clearly and consistently identified within and between populations. This

303 suggests that there are two reproducible HGSC subtypes that are either etiologically distinct, or

304 acquire phenotypically determinant alterations through their development. Our study also

305 suggests that the previously described "immunoreactive-like" and "differentiated-like" subtypes

306 appear more variable across populations, and tend to be collapsed into a single category when

307 three subtypes are specified. These may represent, for example, steps along an immunoreactive

308 continuum or could represent the basis of a third, but more variable subtype.

309 Our analysis also reveals the importance of critically reassessing molecular subtypes

310 across multiple large study populations using parallel analyses and consistent inclusion criteria.

311 New systematic approaches hold promise for the implementation of such analyses (Celik *et al.*

312 2016; Planey and Gevaert 2016). Our results underscore the importance of ovarian cancer

313 histopathology, contradict the four HGSC subtype hypothesis, and suggest that there may be

314 fewer HGSC molecular subtypes with variable immunoreactivity and stromal infiltration.

315

316 <u>ACKNOWLEDGEMENTS:</u>

319

328

329   FIGURE LEGENDS:

330   **Figure 1.** Significance analysis of microarray (SAM) moderated t score Pearson correlation

331   heatmaps reveal consistency across datasets. (A) Correlations across datasets for k means $k = 2$.

332   (B) Correlations across datasets for k means $k = 3$. (C) Correlations across datasets for k means k

333    $= 4$

334

335   **Figure 2.** Significance analysis of microarray (SAM) moderated t score Pearson correlation

336   heatmaps of clusters formed by k means clustering and NMF clustering reveals consistency

337   across clustering methods. Within dataset results are shown for both methods when setting each

338   algorithm to find 2, 3, and 4 clusters.

339

340   **Figure 3.** Comparing NMF consensus clustering in the Tothill dataset. Data displays consensus

341   clustering for $k = 2$ to $k = 6$ for 10 NMF initializations alongside the cophenetic correlation

342   results for $k = 2$ to $k = 8$. (A) Tothill dataset (n = 260) with low malignant potential (LMP)

15

343   samples (n = 18) not removed prior to clustering. (B) Tothill dataset with LMP samples removed

344   (n = 242).

345

346   **Supplementary Figure S1.** Overlapping genes assayed using either the HG-U1133 Affymetrix

347   platform (TCGA, Tothill, Bonome) or the Agilent 4x44K platform (Mayo, Yoshihara).

348   Differences across datasets arise from inherent array differences and/or differences in quality

349   control preprocessing.

350

351   **Supplementary Figure S2.** Sample by sample Pearson correlation matrices. Top panel: $k = 2$.

352   Middle panel: $k = 3$. Bottom panel: $k = 4$. The color bars are coded as blue for cluster 1, red for

353   cluster 2, green for cluster 3, and purple for cluster 4. In the matrices, red represents high

354   correlation, blue low correlation, and white intermediate correlation. The scales are slightly

355   different in each population because of different correlational structures. The clusters in the

356   Bonome study are depicted in gray scale because in cross-population analyses to identify

357   analogous clusters, those from Bonome did not correlate with those observed in the four other

358   studies.

359

360   **Supplementary Figure S3.** NMF consensus matrices for datasets when $k = 2$, $k = 3$, and $k = 4$.

361   The first track represents cluster membership for k means clusters and the second track

362   represents silhouette widths. Note that NMF clusters are not ordered in the same way as the k

363   means clusters.

364

16

365   **Supplementary Figure S4.** Significance analysis of microarray (SAM) moderated t score

366   Pearson correlation heatmaps are not consistent across datasets for randomly shuffled gene

367   expression values for $k = 2$, $k = 3$, or $k = 4$. The within dataset correlations are artificially

368   induced because the clustering algorithm will find clusters even without true underlying

369   structure. However, the across dataset clusters are not correlated in the randomized data

370   indicating that the results we observe in Figure 1 are not artifacts of the clustering algorithm.

371

372   **Supplementary Figure S5.** Consensus NMF clustering of the TCGA dataset (n = 499) for $k = 2$

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

374

375   **Supplementary Figure S6.** Consensus NMF clustering of the Mayo dataset (n = 379 for $k = 2$ to

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

377

378   **Supplementary Figure S7.** Consensus NMF clustering of the Yoshihara dataset (n = 256) for k

379    $= 2$ to $k = 6$ for 10 NMF runs alongside the cophenetic correlation results for $k = 2$ to $k = 8$.

380

381   **Supplementary Figure S8.** Silhouette width plots for $k = 2$, $k = 3$, and $k = 4$ for k means

382   clustering results. Cluster 1 is shown in blue, cluster 2 in red, cluster 3 in green, and cluster 4 in

383   purple.

384

385   **Supplementary Figure S9.** Kaplan-Meier survival curves for $k = 2$, $k = 3$, and $k = 4$ shown for

386   clustering solutions using k means and NMF. Cluster 1 is shown in blue, cluster 2 in red, cluster

387   3 in green, and cluster 4 in purple.

17

388

389   REFERENCES:

390   Blagden, S. P., 2015 Harnessing Pandemonium: The Clinical Implications of Tumor

391   Heterogeneity in Ovarian Cancer. Front. Oncol. 5.:

392   Boettiger, C., 2015 An introduction to Docker for reproducible research. ACM SIGOPS Oper.

393   Syst. Rev. 49: 71–79.

394   Bonome, T., J.-Y. Lee, D.-C. Park, M. Radonovich, C. Pise-Masison *et al.*, 2005 Expression

395   profiling of serous low malignant potential, low-grade, and high-grade tumors of the

396   ovary. Cancer Res. 65: 10602–10612.

397   Bonome, T., D. A. Levine, J. Shih, M. Randonovich, C. A. Pise-Masison *et al.*, 2008 A gene

398   signature predicting for survival in suboptimally debulked patients with ovarian cancer.

399   Cancer Res. 68: 5478–5486.

400   Broad Institute TCGA Genome Data Analysis Center, 2014 Analysis Overview for Ovarian

401   Serous Cystadenocarcinoma (Primary solid tumor cohort) - 15 July 2014.

402   Broad Institute TCGA Genome Data Analysis Center, 2015 Clustering of mRNA expression:

403   consensus NMF.

404   Brunet, J.-P., P. Tamayo, T. R. Golub, and J. P. Mesirov, 2004 Metagenes and molecular pattern

405   discovery using matrix factorization. Proc. Natl. Acad. Sci. 101: 4164–4169.

406   Celik, S., B. A. Logsdon, S. Battle, C. W. Drescher, M. Rendi *et al.*, 2016 Extracting a low-

407   dimensional description of multiple gene expression datasets reveals a potential driver for

408   tumor-associated stroma in ovarian cancer. Genome Med. 8.:

409   Edgar, R., M. Domrachev, and A. E. Lash, 2002 Gene Expression Omnibus: NCBI gene

410   expression and hybridization array data repository. Nucleic Acids Res. 30: 207–210.

411   Ganzfried, B. F., M. Riester, B. Haibe-Kains, T. Risch, S. Tyekucheva *et al.*, 2013

412   curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome.

413   Database 2013: bat013–bat013.

414   Gregory Way, James Rudd, and Casey Greene, 2015 Analytical Code for "Cross-population

415   analysis of high-grade serous ovarian cancer reveals only two robust subtypes."

416   Konecny, G. E., C. Wang, H. Hamidi, B. Winterhoff, K. R. Kalli *et al.*, 2014 Prognostic and

417   Therapeutic Relevance of Molecular Subtypes in High-Grade Serous Ovarian Cancer.

418   JNCI J. Natl. Cancer Inst. 106: dju249–dju249.

419   Kurman, R. J., and I.-M. Shih, 2010 The Origin and Pathogenesis of Epithelial Ovarian Cancer:

420   A Proposed Unifying Theory: Am. J. Surg. Pathol. 34: 433–443.

421   Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, 2014 cluster: Cluster

422   Analysis Basics and Extensions. R package version 1.15.3.:

423   Ouellet, V., D. M. Provencher, C. M. Maugard, C. Le Page, F. Ren *et al.*, 2005 Discrimination

424   between serous low malignant potential and invasive epithelial ovarian tumors using

425   molecular profiling. Oncogene 24: 4672–4687.

426   Planey, C. R., and O. Gevaert, 2016 CoINcIDE: A framework for discovery of patient subtypes

427   across multiple datasets. Genome Med. 8.:

428   Schwender, H., 2012 siggenes: Multiple testing using SAM and Efron's empirical Bayes

429   approaches. R package version 1.40.0.:

430   Schwender, H., A. Krause, and K. Ickstadt, 2006 Identifying interesting genes with sigenes.

431   RNews 6: 45–50.

432   Silverberg, S. G., 2000 Histopathologic grading of ovarian carcinoma: a review and proposal.

433   Int. J. Gynecol. Pathol. Off. J. Int. Soc. Gynecol. Pathol. 19: 7–15.

434   Sokal, R. R., and F. J. Rohlf, 1962 The Comparison of Dendrograms by Objective Methods.

435   Taxon 11: 33.

436   Soslow, R. A., 2008 Histologic Subtypes of Ovarian Carcinoma: An Overview. Int. J. Gynecol.

437   Pathol. PAP.:

438   Tan, T. Z., Q. H. Miow, R. Y.-J. Huang, M. K. Wong, J. Ye *et al.*, 2013 Functional genomics

439   identifies five distinct molecular subtypes with clinical relevance and pathways for

440   growth control in epithelial ovarian cancer: A subtyping scheme for epithelial ovarian

441   cancer. EMBO Mol. Med. 5: 1051–1066.

442   The Cancer Genome Atlas, 2011 Integrated genomic analyses of ovarian carcinoma. Nature 474:

443   609–615.

444   Tothill, R. W., A. V. Tinker, J. George, R. Brown, S. B. Fox *et al.*, 2008 Novel Molecular

445   Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome. Clin.

446   Cancer Res. 14: 5198–5208.

447   Tusher, V. G., R. Tibshirani, and G. Chu, 2001 Significance analysis of microarrays applied to

448   the ionizing radiation response. Proc. Natl. Acad. Sci. 98: 5116–5121.

449   Vang, R., I.-M. Shih, and R. J. Kurman, 2009 Ovarian Low-grade and High-grade Serous

450   Carcinoma: Pathogenesis, Clinicopathologic and Molecular Biologic Features, and

451   Diagnostic Problems. Adv. Anat. Pathol. 16: 267–282.

452   Verhaak, R. G. W., P. Tamayo, J.-Y. Yang, D. Hubbard, H. Zhang *et al.*, 2012 Prognostically

453   relevant gene signatures of high-grade serous ovarian carcinoma. J. Clin. Invest.

454   Yoshihara, K., T. Tsunoda, D. Shigemizu, H. Fujiwara, M. Hatae *et al.*, 2012 High-Risk Ovarian

455   Cancer Based on 126-Gene Expression Signature Is Uniquely Characterized by

456   Downregulation of Antigen Presentation Pathway. Clin. Cancer Res. 18: 1374–1385.

457

458

459

460

461

462

463

464

465

466   **Table 1**: Characteristics of the populations included in the five 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%)

467

468   NA: Data not reported

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

470   [b]samples without survival data were excluded in survival analyses

471   **Table 2**: SAM moderated *t* score vector Pearson correlations between analogous clusters across

472   populations[a]

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
$k = 2$[a]	0.62 – 0.81	0.62 – 0.81	NA	NA
$k = 3$[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 $k = 2$[b]	-0.08 – 0.24	-0.08 – 0.24	NA	NA
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

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

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

475

476

477

478

479

480

481

482

483

23

484   **Table 3:** Distributions of sample membership in the clusters identified in our study by the

485   original cluster assignments in the TCGA, Tothill, and Konecny studies. Clusters identified in

486   our study using *k*-means clustering with $k = 2$, $k = 3$, and $k = 4$

	TCGA					Tothill *et al.*							Konecny *et al.*				
$k = 2$	Mes	Pro	Imm	Dif	NC[a]	C1	C2	C3	C4	C5	C6	NC[a]	C1	C2	C3	C4	NA[b]
Cluster 1	98	7	93	68	21	78	39	1	0	0	0	11	36	21	2	26	114
Cluster 2	1	127	2	60	22	0	5	5	44	35	2	22	6	39	41	0	94
$k = 3$																	
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
$k = 4$																	
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

487

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

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

490   NOTE: The corresponding labels for the generally similar HGSC gene expression subtypes

491   observed in the TCGA, Tothill, and Konecny studies are, respectively: mesenchymal/C1/C4,

492   proliferative/C5/C3, immunoreactive/C2/C1, and differentiated/C4/C2)

24

A.

B.

	TCGA	Mayo	Yoshihara	Tothill	Bonome
k = 2					
k = 3					
k = 4					
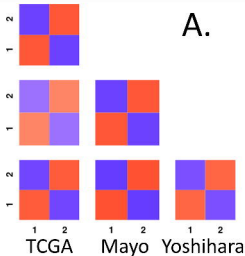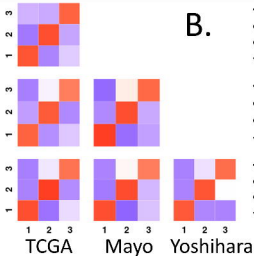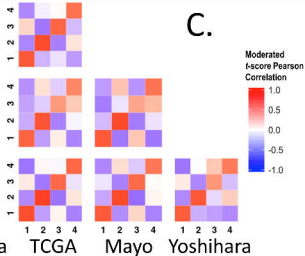
TCGA

Mayo

Yoshihara

TCGA

1 : 287 | 0.10

2 : 212 | 0.008

1 : 137 | 0.10

2 : 139 | 0.01

3 : 223 | 0.02

1 : 130 | 0.09

2 : 99 | 0.01

3 : 100 | 0.03

4 : 170 | 0.01

Yoshihara

1 : 97 | 0.13

2 : 159 | 0.03

1 : 92 | 0.12

2 : 44 | 0.003

3 : 120 | 0.05

1 : 88 | 0.11

2 : 30 | 0.03

3 : 74 | 0.04

4 : 64 | 0.04

Mayo

1 : 199 | 0.14

2 : 180 | 0.01

1 : 139 | 0.10

2 : 108 | 0.008

3 : 132 | 0.04

1 : 109 | 0.07

2 : 81 | 0.03

3 : 93 | 0.05

4 : 96 | 0.01

Tothill

1 : 129 | 0.16

2 : 113 | 0.02

1 : 105 | 0.14

2 : 46 | 0.04

3 : 91 | 0.04

1 : 73 | 0.14

2 : 43 | 0.03

3 : 67 | 0.02

4 : 59 | 0.01
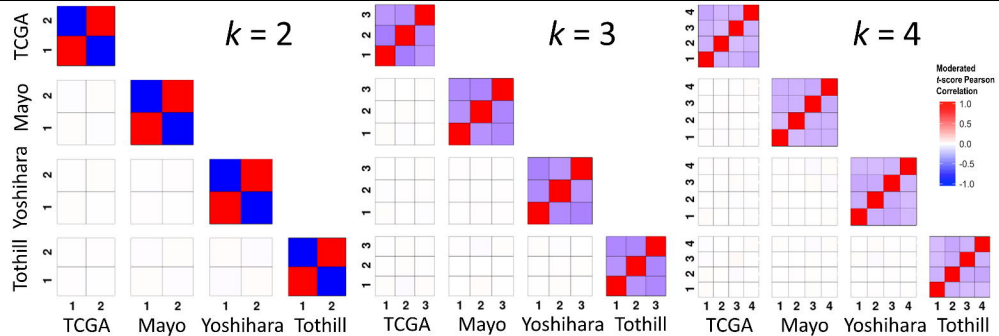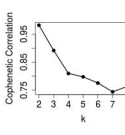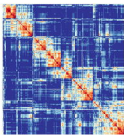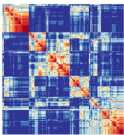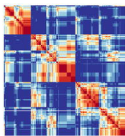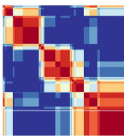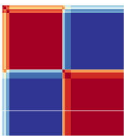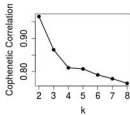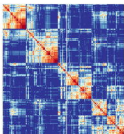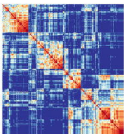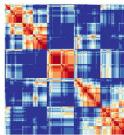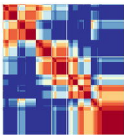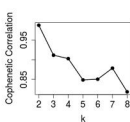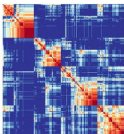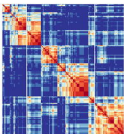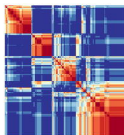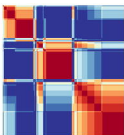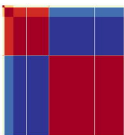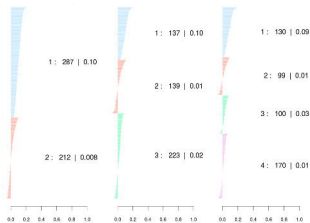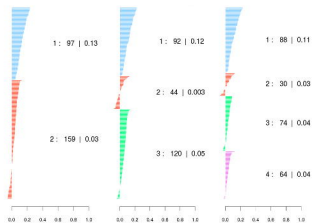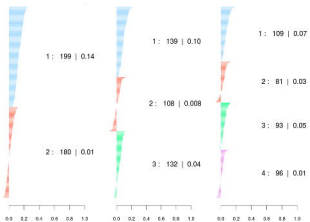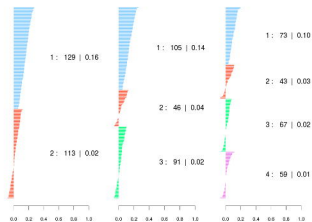
TCGA — Yoshihara — Mayo — Tothill survival curves. Cumulative Survival vs Months for *k means* and *NMF* clustering methods at *k* = 2, 3, 4. Clusters: 1 (blue), 2 (orange/red), 3 (green), 4 (magenta).