

22 **Abstract**

23 To stratify cancer patients for most beneficial therapies, it is a priority to define robust
24 molecular subtypes using clustering methods and “big data”. If each of these methods produces
25 different numbers of clusters for the same data, it is difficult to achieve an optimal solution. Here,
26 we introduce “polyCluster”, a tool that reconciles clusters identified by different methods into
27 context-specific subtype “communities” using a hypergeometric test or a measure of relative
28 proportion of common samples. The polycluster was tested using a breast cancer dataset, and latter
29 using uveal melanoma datasets to identify novel subtype communities with significant metastasis-
30 free prognostic differences. Available at: <https://github.com/syspremed/polyClustR>

31

32

33

34

35

36

37

38

39

40 **Background**

41 Recently, advances in omics technologies have lead to large volumes of data being collected
42 on molecular profiles, including gene expression, in various cancers. Cancers of all types exhibit
43 inter-tumoral (between patient) heterogeneity that can be quantified in part by gene expression. This
44 heterogeneity can help explain the differential prognosis in cancer patients treated with the same
45 therapies. A well-established example is the specific efficacy of trastuzumab (Herceptin) in HER2-
46 positive breast cancer [1]. Previously, we have suggested potential differential cetuximab (anti-
47 EGFR therapy) response in colorectal cancer (CRC) subtypes [2]. More recently, trials of
48 oxaliplatin in Stage II and III CRC found that its effectiveness may be limited to certain subtypes
49 published by us [2, 3]. In pancreatic cancer, we observed a relatively increased response to
50 gemcitabine in quasi-mesenchymal (QM) subtype cell lines compared to classical subtype cell lines
51 [4]. This result corroborates with the finding by Mofitt *et al.*, that patients from basal-like pancreatic
52 cancer subtype (equivalent to our QM subtype) has improved response to adjuvant therapy compared
53 to classical subtype pancreatic tumors [5]. Similarly, we showed potential subtype-specific
54 therapies using a panel of breast cancer cell lines and drug response analysis [6]. Nevertheless, for
55 accurate prediction of therapy responses, the challenge lies in defining robust and clinically relevant
56 subtypes.

57

58 In breast cancer, where current opinion lies with the existence of 5 intrinsic gene expression
59 subtypes (basal, HER2/ERBB2, luminal A, luminal B, and normal-like), studies have variously
60 reported a number of subtypes ranging between 4 [7] and 10 [8]. While multiple factors are
61 involved in this apparent discrepancy in defining a number of cancer subtypes, the clustering
62 methodologies can also significantly contribute to this difference. There are various clustering
63 algorithms that are regularly employed for this purpose, and each has its own strengths according to

64 the underlying structure of the data it is applied to. As clustering algorithms have a huge range of
65 potential applications, selection of the appropriate algorithm to use in any given situation can be
66 difficult. At the same time, the need for the user to inspect the results of each algorithm over a range
67 of numbers of clusters (k) and select the optimal solution are often subjective. This situation has
68 been improved by the adoption of various consensus clustering techniques, which allow for visual
69 and quantitative examination of multiple re-runs of the same algorithm so the effects of random
70 starting points can be taken into consideration.

71 However, consensus clustering does not diminish the influence the choice of algorithm has
72 on the clustering solution. The application of different consensus clustering algorithms leads to
73 different number of subtypes (number of clusters, k), and hence, defining the optimal number of
74 clusters is often challenging. This is due to various factors in the design of the algorithm: whether it
75 is ‘greedy’, that is, if it makes the locally optimal choice at each individual stage at the possible
76 expense of finding a global optimum; whether cluster centroids must be located at data points; and
77 how iterative algorithms evaluate their convergence to a solution are some examples [9]. This
78 makes the use of a single algorithm to cluster gene expression profiles, as is often done in subtyping
79 studies, risky. In addition, the clusters found may well be valid, but information about either larger
80 stratification of the data or small but distinct sub-subtypes of low frequency may be lost [10]. It is
81 for this reason that finding methods of reconciling optimal clustering solutions identified by
82 different algorithms is necessary. Cluster reconciliation not only validates the clusters from
83 different algorithms – it can also reveal in greater detail the structure in the data on the macro and
84 the micro scale, from broad classifications resulting from a handful of important functional groups,
85 to rarer and less well-defined sub-subtypes. It also reveals more about the efficacy of the clustering
86 algorithms themselves [10, 11].

87 Here, we demonstrate how to identify optimal solutions and define subtype “communities”
88 by reconciling clusters identified from three different consensus clustering methods - hierarchical

89 clustering (HC) [12, 13], k-means (KM) [14], and non-negative matrix factorization (NMF) [15].
90 The clusters were further reconciled using at least two approaches. The first, a hypergeometric test
91 to determine the probability that two clusters share the same samples by chance, was previously
92 used to successfully reconcile subtypes of CRC found via clustering in two studies which found
93 three and five optimal subtypes, respectively [2, 10, 16]. It was determined via this analysis that the
94 three subtypes could be appropriately divided into the five sub-subtypes. When four further studies
95 into CRC were published, finding between 3 and 6 optimal clusters [17-20], the Jaccard index was
96 applied to help understand the relationships between these solutions and find “consensus molecular
97 subtypes” (CMS) [11]. The second and a new reconciliation measure used here – calculating the
98 relative proportion of samples in a smaller cluster present in a larger one (termed Eason-
99 Sadanandam index) – differs from measures of cluster similarity such as the Jaccard index in order
100 to give sub-subtypes a high score, even if they are much smaller than a larger cluster (see **Methods**
101 section).

102 All the above reconciliation methods are part of our new framework or package called
103 “polyCluster”. The framework is flexible that other methods can be included any time. Here, we
104 demonstrate how our new framework can be used to identify breast cancer gene expression
105 “subtype communities” and to compare with existing intrinsic subtypes [7]. Moreover, we have
106 applied this to uveal melanoma gene expression profiles to define novel gene expression “subtype
107 communities” with different prognosis and chromosomal aberrations associated with them.

108

109

110

111

112

113

114

115 **Results and Discussion**

116 Our reconciliation method (**Figure 1**) uses a matrix of preprocessed and normalized gene
117 expression (or any other similar data) and performs the following: a) applies different consensus
118 clustering methods (including NMF, HC and KM) and uses statistical scores (specific to each
119 method described below) for each clustering to determine the optimal number of clusters; and b)
120 reconciles the results from different clustering methods and identifies a consensus solution by
121 creating network of clusters that defines communities of integrated subtypes using methods such as
122 the hypergeometric test and proportion of maximum intersection (PMI). We then identify the
123 optimal “community” with highest average silhouette width [21] and compare this reconciliation to
124 known subtypes, if they exist, for that set of samples. To illustrate this, we used published gene
125 expression profiles from breast cancer and uveal melanoma as examples.

126

127 **Application to reconcile breast cancer “subtype communities” with intrinsic subtypes**

128 *Breast cancer subtypes defined by multiple clustering methods*

129 For this purpose, we used breast tumor gene expression data ($n = 118$) from a published
130 study [22]. Details of initial clustering of this dataset and selection of k clusters for each algorithm
131 are provided in **Figures S1A-C**. Initially, we applied the NMF to the 2258 most highly variable
132 genes from this Chin data set as selected by standard deviation ($SD > 0.8$). We identified highest
133 cophenetic correlation coefficient of 0.9997 at k subtypes for NMF $k_{NMF}=2$ followed by 0.9962 at
134 $k_{NMF}=6$. Silhouette width also showed peaks at k_{NMF} at 2 and 6 (**Figures S1 A-C**). In order to
135 capture the most heterogeneity, we chose $k_{NMF}=6$, and named the clusters breast cancer (b)NMF1 to
136 6. Overall, known subtypes of these samples [22] were significantly associated with these clusters
137 (Fisher’s exact test; $p < 0.001$). Specifically, the clusters bNMF1, bNMF3 and bNMF4 were
138 significantly associated with luminal A, basal and luminal B, respectively (hypergeometric test;
139 false discovery rate; $FDR < 0.01$) (**Figure 2A**). The basal subtype was also border-line significantly

140 associated ($FDR=0.2$) with bNMF5, suggesting the existence of a sub-subtype of basal breast
141 cancer that was not identified earlier when subtypes for this dataset were predicted by correlation
142 with intrinsic subtype signatures [23] [7]. bNMF2 and bNMF6 were not significantly associated
143 with any of the published subtypes. Gene set enrichment analysis (GSEA) of these unidentified
144 subtypes revealed associations with metaplastic breast cancer (bNMF2, $FDR<0.01$) and with
145 17q21-q25 amplicon gene sets (bNMF6, $FDR < 0.1$) (**Figure S2A-B**). Overall, application of NMF
146 to the Chin data set identified clusters that partially overlapped with published subtypes, and others
147 with interesting breast cancer biology.

148

149 Since NMF identified extra subtypes in Chin data set, we applied two additional clustering methods
150 – consensus hierarchical clustering (HC) and K-means (KM). When we applied consensus
151 hierarchical clustering to the same data, $k_{HC}=2$ and $k_{HC}=6$ had the highest silhouette widths.
152 (**Figures S1A and C**). The cophenetic coefficient after $k_{HC}=6$ does not increase significantly and
153 the consensus plot showed consensus clusters (**Figures S1A and C**). Hence, we chose six HC
154 clusters to again cover the most heterogeneity. The clusters from HC for breast cancer data were
155 defined as breast cancer (b)HC. As with the NMF clusters, these clusters were significantly
156 associated with the known subtypes of these samples (Fisher's exact test; $FDR<0.001$). The bHC1,
157 bHC3 and bHC6 clusters were significantly (hypergeometric test; $FDR<0.01$) associated with basal,
158 luminal A and normal-like subtypes, respectively (**Figure 2B**). Both bHC2 and bHC5 were
159 significantly ($FDR<0.01$) associated with luminal B. bHC4 was marginally significantly associated
160 with luminal A subtype, and bHC5 with the ERBB2 (HER2) subtype, with less significance
161 ($FDR<0.2$; **Figure 2B**).

162

163 Additionally, we applied consensus KM clustering to the Chin data set. While both the
164 cophenetic coefficient and silhouette width showed highest peaks at $k_{KM}=3$ and 4 (after $k_{KM}=2$), we

165 observed that consensus clustering at these k_{KM} s did not show clear consensus clusters. There were
166 not large differences in cophenetic coefficient, silhouette width and consensus clusters at k_{KM}
167 between 4 and 7 (**Figure S1A and D**). Hence, we chose $k_{KM}=7$ as an optimal cluster. All of these
168 KM clusters (defined as breast cancer (b)KM) were significantly associated with known breast
169 cancer subtypes (**Figure 2C**; Fisher's exact test; $p < 0.001$), unlike the NMF and HC clusters.
170 Specifically, bKM1 and bKM4 were associated with basal, bKM2 with luminal B and bKM3,
171 bKM5 and bKM6 with luminal A (hypergeometric test; $FDR < 0.01$). bKM7 was significantly
172 associated with the ERBB2 subtype, which was not highly significant with any NMF or HC
173 clusters. bKM3 was marginally associated with the normal-like subtype ($FDR=0.08$). Direct
174 comparison of the two basal clusters through GSEA revealed enrichment of multiple gene sets
175 associated with invasive breast cancer, immunity and cytokines (**Figure S2C-F**). This clearly
176 suggests that different clustering algorithms have the inherent capacity to identify distinct clusters.
177 Here, KM has identified clusters with more significant association to published subtypes.

178

179 *Identification of breast cancer "subtype communities"*

180 The existence of multiple clustering solutions defined by different algorithms poses the
181 question of what number of clusters is optimal, and how they reconcile between different methods.
182 To address these questions, we chose two different reconciliation methods – hypergeometric test
183 and proportion of maximum intersection. The results from each of the reconciliation methods are
184 discussed below.

185

186 Previously, we have used the hypergeometric test to assess enrichment of samples between
187 two CRC classifications (including ours) as a means of reconciling subtypes [10]. Similarly, we
188 have used this analysis here to reconcile breast cancer clusters between the three different (NMF,
189 HC and KM) algorithms utilized above. Subsequently, in order to group those clusters with

190 significant similarity into “subtype communities” , we performed network community detection by
191 applying weighted label propagation method (using FDR values as edge weights) [24]. As a result,
192 we observed six “subtype communities” (groups of clusters; bHYP1-6) based on this analysis
193 (**Figure 3A**).

194

195 There was significant association with the known subtypes and these communities (Fisher’s
196 exact test; $p < 0.001$). We observed that five communities were primarily and significantly
197 (hypergeometric test; $FDR < 0.05$) associated with published breast cancer subtypes – bHYP3 and
198 bHYP4 with luminal A, bHYP2 with luminal B and bHYP1 and bHYP6 with basal (**Figures 3A**
199 **and S3A**). Four of the communities (bHYP1-4) contained clusters from all three clustering
200 algorithms (**Figure 3A**). Interestingly, each of the luminal A and basal subtypes were split into two
201 communities. One basal community (bHYP6) contained the immune-enriched bKM4 cluster. One
202 of the luminal A communities (bHYP3) contained a number of samples from the ERBB2 subtype in
203 a cluster that was enriched for a metaplastic breast cancer signature (bNMF2; **Figures 3A** and
204 **S3A**), while the other (bHYP4) contained some luminal B samples in the 17q21-q25 amplicon-
205 enriched cluster (bNMF6; **Figures 3A** and **S3A**). Finally, there was a community (bHYP5) with
206 mixture of normal-like and ERBB2 subtype samples. This community was the most mixed in terms
207 of intrinsic subtypes. Overall, hypergeometric test-based reconciliation expanded the breast cancer
208 subtypes to 6 communities.

209

210 Our PMI method is similar to the Jaccard analysis that we used recently to reconcile CRC
211 subtypes as a part of the CRC Subtyping Consortium (CRCSC) [11], with the difference that it
212 weights sub-groups of a larger cluster as strongly as identical clusters of the same size (see
213 **Methods**). Here, we applied the PMI method to reconcile subtypes from NMF, HC and KM similar
214 to what we performed using the hypergeometric test. Unlike the hypergeometric method, PMI

215 identified five communities (bPMI1 to 5; **Figures 3B** and **S3B**), four (bPMI2 to 5) of which were
216 analogous to hypergeometric communities (bHYP2, 3, 4 and 5). The final community (bPMI1) was
217 a combination of the two basal hypergeometric communities (bHYP1 and 6). These communities
218 were significantly associated with known subtypes, overall (Fisher’s exact test; $p < 0.001$). As
219 expected, four of the five communities represent luminal A (bPMI3 and 4), luminal B (bPMI2) and
220 basal (bPMI1) communities (hypergeometric; $FDR < 0.05$). The other community (bPMI5) was a
221 mixture of HER2/ERBB2 and normal-like (**Figures 3B** and **S3C**).

222

223 To chose optimal “subtype community” between HYP and PMI communities, we calculated
224 the silhouette width [21] for all samples in the different communities (**Figures 3** and **S4**). The
225 average silhouette widths for HYP communities were 0.06 and that for PMI communities were
226 0.07. Hence, PMI communities with highest average silhouette width were chosen as optimal.

227

228 This application of the pipeline to a well-characterised cancer has demonstrated its ability to
229 identify new biologically distinct “subtype communities” of patients, alongside those subtypes
230 which have already been extensively described. We next sought to apply this pipeline to a cancer
231 with molecular subtypes that have not been explored so comprehensively, although uveal melanoma
232 classes at gene expression levels are known [25-27].

233

234 **Application to uveal melanoma and identification of novel “subtype communities”**

235 *Identification of subtype communities*

236 Compared to breast cancer, uveal melanoma is a cancer type that has not been extensively
237 subtyped, presumably due to its low incidence. This scarcity of samples makes clustering a
238 challenge – clusters discovered are less likely to be robust due to their small size. It is in cases such

239 as this where the reconciliation of clusters from multiple algorithms may present benefits in terms
240 of increasing confidence in the results of clustering.

241

242 As with the breast cancer data, we applied the three clustering algorithms of HC, KM and
243 NMF to a dataset of the 6146 most variable genes ($SD > 0.8$) from 58 patients with uveal melanoma
244 (GSE22138, [28]). By performing the same assessment of cophenetic coefficient, silhouette width
245 and consensus matrices, we discovered four clusters by HC, six clusters by KM and five clusters by
246 NMF (**Figure S5A-D**). This demonstrates that different clustering methods yield different clusters
247 using the same data set. However, reconciling the results from these methods to identify the optimal
248 number of clusters can characterize the more heterogeneity in uveal melanoma that may be
249 associated with disease phenotypes such as metastasis and abnormalities in chromosome 3.

250

251 By reconciling these subtypes by a hypergeometric test followed by community detection,
252 we identified five “subtype communities” of clusters (**Figure 4A**). When we assessed these
253 communities for the key molecular feature of chromosome 3 aneuploidy, we discovered a
254 significant association of these communities with this feature (Fisher’s exact test; $p < 0.001$); one
255 community – melanoma mHYP2 – was significantly enriched (hypergeometric test; $FDR < 0.001$)
256 for monosomy, and another (mHYP5) was significantly enriched ($FDR < 0.05$) for both disomy and
257 partial monosomy (**Figures 4A** and **S6A**). Two of the remaining three communities showed less
258 significant associations with chromosome 3 disomy (mHYP4) and monosomy (mHYP1;
259 hypergeometric test; $FDR < 0.2$) respectively, while the final community (mHYP3) was not
260 significantly enriched for either. A similar pattern of associations was observed when assessing four
261 “subtype communities” defined by the PMI method (**Figure 4B**), with one community each
262 representing monosomy and disomy (mPMI1 and mPMI4, respectively), and one mixed
263 disomy/partial monosomy/monosomy community (mPMI2) – however the association was not

264 statistically significant (Fisher's exact test; $p=0.577$). (**Figures 4B** and **S6B**). HYP subtypes were
265 chosen over PMI subtypes for significant association with known key molecular features of uveal
266 melanoma and having lower number of samples with negative silhouette width in this cohort
267 (**Figure S7**).

268

269 *Biological understanding of uveal melanoma subtype communities*

270 Next, we sought to understand these communities by performing GSEA, and discovered that
271 one of these communities (mHYP1) was significantly enriched ($FDR<0.05$) for gene sets associated
272 with immune pathways (e.g. cytokine-cytokine receptor interactions, T cell receptor signaling and
273 JAK-STAT pathway; $FDR<0.05$; **Figure 5A-D**). On the other hand, another subtype (mHYP3) was
274 associated with neural cell types (e.g. neuron markers, neurotransmitter signaling, neural subtype
275 glioblastoma; **Figure 5E-H**; $FDR<0.05$). The last communities (mHYP2, mHYP4 and mHYP5) did
276 not significantly associate with any gene sets. This could indicate that mHYP2 enriched for
277 chromosome 3 monosomy and mHYP4 may be by disomy, may be defined by that particular
278 phenotype as opposed to a coherent transcriptomic pattern.

279

280 *Patient prognostic differences between uveal melanoma subtype communities*

281 Since more than 50% uveal melanoma patients undergo metastasis [28], we assessed the
282 metastasis-free prognosis of the uveal melanoma subtype communities using the GSE22138 [28]
283 data set. Among the two highly frequent communities, mHYP2 (37%) showed significantly poorest
284 metastasis-free prognosis, whereas mHYP5 (28%) showed better prognosis. Both mHYP4 (20%)
285 and mHYP1 (11%) communities showed intermediate prognosis (**Figure 6A**).

286

287 *Validation of uveal melanoma subtype communities*

288 Due to the low frequency of some of these communities in this dataset (5% mHYP3, 11%
289 mHYP1), we sought to validate them in an independent dataset consisting of 58 patients with uveal
290 melanoma (GSE44295). Patients were assigned to subtypes based on the correlation of their gene
291 expression profile with the prediction analysis of microarrays (PAM) [29] centroids of each
292 community. In the validation cohort, 31% of patients were assigned to the mHYP1 (immune-
293 enriched) group, 19% mHYP2 (monosomy-enriched), 14% mHYP3 (neural-enriched), 5% mHYP4
294 (undetermined) and 31% mHYP5 (disomy/partial monosomy-enriched). In terms of prognosis,
295 these groups showed statistically significant differential metastasis-free survival ($p = 0.00747$;
296 **Figure 6B**). Analogous to the previous dataset, mHYP2 and mHYP5 communities showed poor and
297 good prognosis respectively. While mHYP1 showed intermediate prognosis, mHYP4 couldn't be
298 assessed due to low sample size of only 5% ($n=3$). Interestingly and similar to the training
299 (GSE22138) dataset, 82% of mHYP2 (monosomy-enriched) group in the validation cohort
300 underwent metastasis during follow-up, compared to only 11% of the mHYP5 (disomy/partial
301 monosomy-enriched) group patients. In addition, 33% of intermediate prognostic mHYP4
302 (undetermined) and 44% mixed prognostic mHYP1 (immune-enriched) patients experienced
303 metastasis. With increased frequency of mHYP3 (neural-enriched) community, we observed that it
304 has poor overall survival and 57% of the mHYP3 samples were undergoing metastasis (**Figure 6B**).
305 Overall, this identifies and validates novel uveal melanoma subtype communities and their
306 prognostic significance.

307

308 *Comparison of subtype communities to known uveal melanoma classes*

309 Previously, transcriptomic subtypes of uveal melanoma have been defined by clustering of
310 gene expression profiles. Two classes were discovered – class 1, with good prognosis and
311 association with chromosome 3 disomy; and class 2, with poor prognosis, associated with
312 chromosome 3 monosomy and metastasis [25-27]. To reconcile these communities with the gene

313 expression subtypes, we checked for gene set enrichment of the gene signatures [26] for class 1 and
314 class 2 uveal melanomas in this cohort. The class 2 signature was enriched and borderline enriched
315 in the mHYP1 community (immune-enriched; $FDR < 0.001$; **Figure 6C**) and mHYP2 (monosomy;
316 $FDR = 0.27$; **Figure 6D**) groups, respectively, whereas, unexpectedly, the class 1 signature was not
317 significantly enriched in any other group. This may indicate that the class 1 signature may be a
318 heterogeneous set of patients who are not confined to any of our given community. Overall, this
319 suggest that our novel uveal melanoma subtype communities reveal additional heterogeneity with
320 clinical significance that requires further investigation.

321

322 **Conclusions**

323 These results demonstrate that no one clustering algorithm should be relied on to produce
324 clusters which are robust and capture all heterogeneity in a dataset. Instead, multiple algorithms
325 should be applied to the same dataset, and their results compared and reconciled. Our polyCluster
326 tool provides a straightforward interface to cluster datasets using multiple algorithms, provides
327 statistics on the quality of each clustering, and allows the user to fully understand how each result is
328 related through multiple reconciliations. The demonstration that some low-frequency clusters –
329 which may be lost or discarded as outliers if only one algorithm is applied – are consistently
330 identified across algorithms lends credence to their validity, and here such communities were
331 additionally validated in an independent dataset. Thus, the reconciliation of multiple clustering
332 results enables finer stratification of patients' molecular profiles enabling more focused biological
333 profiling.

334

335

336 **Methods**

337 **Datasets**

338 The breast cancer dataset [22] consists of 118 gene expression profiles generated from
339 frozen resected samples. Patients in this were mostly early-stage, and were a mixture of node- and
340 ER-positive and -negative. The discovery uveal melanoma dataset (GSE22138 [28]) consists of
341 gene expression profiles for 63 untreated patients, chromosome 3 monosomy status and follow-up
342 metastasis-free survival information. The validation dataset (GSE44295 [30]) contains 58 gene
343 expression profiles from enucleation specimens, with metastasis-free survival information.

344

345 **Finding the optimal number of clusters**

346 It is not optimal for each of the above clustering methods to find local solutions which
347 depend on the initial conditions, rather than robust clusterings that are stable over various input
348 parameters. To address this, consensus clustering approaches repeat several iterations of the same
349 algorithm using different random starting points, and can also perform the clustering over different
350 subsets of samples. Consensus clustering for each algorithm was performed over a range of k -values
351 from 2 to 10 and over multiple subsets of the data. The results of the consensus clustering were then
352 inspected in order to determine the optimal k . Determining the optimal k from visual inspection
353 alone is subjective, and so quantification of the consensus clustering is required. Here, the
354 cophenetic correlation coefficient [31] and the silhouette width [21] were used to score each
355 clustering.

356

357 **Hypergeometric test**

358 Previous works have used the hypergeometric test to determine if different algorithms'
359 subtypes correspond to one another [10]. In this pipeline, comparisons can be made between any

360 number of clustering algorithms. The hypergeometric test based false discovery rate (FDR)
361 indicating the significance of the size of the overlap between two clusters was used.

362

363 **Statistical analysis**

364 FDR values for enrichment of gene sets were reported as calculated by the Broad Institute's
365 GSEA software [32]. Kaplan-Meier analysis was used to assess survival and p-values determined
366 from the log-rank test. PAM analysis to generate centroids and assign subtypes using Pearson
367 correlation and gene expression data was done as previously described [11].

368

369 **Software**

370 Code for hierarchical and k-means consensus clustering was adapted from the
371 *ConsensusClusterPlus* v1.36.0 [33] R package. NMF was performed via the *nmf* v0.20.6 R package
372 [34]. The *igraph* R package v1.0.1[35] was used for plotting networks and community detection.
373 Silhouette width was calculated and plotted using the *silhouette* function from the R package *cluster*
374 v2.0.4 [36]. Survival analysis was performed using the *survival* v2.39-5 R package [37]. GSEA was
375 performed using the Broad Institute GSEA software [32]. The pipeline described in this paper is
376 publicly available on GitHub at <https://github.com/syspremed/polyClustR>.

377

378

379

380

381

382 **Declarations**

383 **Availability of data and material**

384 Gene expression data analysed in this study are publicly available from the original publications
385 (breast cancer data [22] and uveal melanoma [28], [30]) and through ArrayExpress with access
386 number E-TABM-158 (breast) and Gene Expression Omnibus (GEO) with accession numbers
387 GSE22138 and GSE44295 (uveal melanoma).

388

389 **Competing interests**

390 A. Sadanandam has ownership interest (including patents) as a patent inventor for a patent entitled
391 "Colorectal cancer classification with different prognosis and personalized therapeutic responses"
392 (patent number PCT/IB2013/060416). No potential conflicts of interest were disclosed by the other
393 authors.

394

395 **Authors' contributions**

396 KE wrote the manuscript, developed the polyCluster package, performed all the experiments and
397 analysed the results. GN helped with the statistical methods and oversaw the data analysis. AS
398 conceived the idea, interpreted the results and wrote the manuscript.

399

400 **Acknowledgments**

401 We acknowledge NHS funding to the NIHR Biomedical Research Centre at The Royal Marsden
402 and the ICR.

403

404

405

406 **References**

- 407 1. Hudis CA: **Trastuzumab — Mechanism of Action and Use in Clinical Practice.** *New*
408 *Engl J Med* 2007, **357**:39-51.
- 409 2. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschlegler S, Ostos
410 LCG, Lannon WA, Grotzinger C, Del Rio M, et al: **A colorectal cancer classification**
411 **system that associates cellular phenotype and responses to therapy.** *Nature*
412 *Medicine* 2013, **19**:619-625.
- 413 3. Song N, Pogue-Geile KL, Gavin PG, Yothers G, Rim Kim S, Johnson NL, Lipchick C,
414 Allegra CJ, Petrelli NJ, O'Connell MJ, et al: **Clinical outcome from oxaliplatin**
415 **treatment in stage II/III colon cancer according to intrinsic subtypes: Secondary**
416 **analysis of NASBP C-07/NRG oncology randomized clinical trial.** *JAMA Oncology*
417 2016, **2**:1162-1169.
- 418 4. Collisson EA, Sadanandam A, Olson P, Gibb WJ, Truitt M, Gu S, Cooc J, Weinkle J, Kim GE,
419 Jakkula L: **Subtypes of pancreatic ductal adenocarcinoma and their differing**
420 **responses to therapy.** *Nature Medicine* 2011, **17**:500-503.
- 421 5. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA, Rashid NU,
422 Williams LA, Eaton SC, Chung AH, et al: **Virtual microdissection identifies distinct**
423 **tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma.**
424 *Nature Genetics* 2015, **47**:1168-1178.
- 425 6. Heiser LM, Sadanandam A, Kuo W-L, Benz SC, Goldstein TC, Ng S, Gibb WJ, Wang NJ,
426 Ziyad S, Tong F, et al: **Subtype and pathway specific responses to anticancer**
427 **compounds in breast cancer.** *Proceedings of the National Academy of Sciences* 2012,
428 **109**:2724-2729.
- 429 7. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT,
430 Johnsen H, Akslen LA, et al: **Molecular portraits of human breast tumours.** *Nature*
431 2000, **406**:747-752.
- 432 8. Wu L, Liu Z, Xu J, Chen M, Fang H, Tong W, Xiao W: **NETBAGs: a network-based**
433 **clustering approach with gene signatures for cancer subtyping analysis.**
434 *Biomarkers in medicine* 2015, **9**:1053-1065.
- 435 9. Han J, Pei J, Kamber M: **Data Mining: Concepts and Techniques.** *Elsevier* 2011, **3rd**
436 **Edition.**
- 437 10. Sadanandam A, Wang X, de Sousa EMF, Gray JW, Vermeulen L, Hanahan D, Medema JP:
438 **Reconciliation of classification systems defining molecular subtypes of colorectal**
439 **cancer: interrelationships and clinical implications.** *Cell Cycle* 2014, **13**:353-357.
- 440 11. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L,
441 Roepman P, Nyamundanda G, Angelino P, et al: **The consensus molecular subtypes**
442 **of colorectal cancer.** *Nature Medicine* 2015, **21**:1350-1356.
- 443 12. Navarro JF, Frenk CS, White SDM: **A Universal density profile from hierarchical**
444 **clustering.** *The Astrophysical Journal* 1997, **490**:493.
- 445 13. Defays D: **An efficient algorithm for a complete link method.** *The Computer Journal*
446 1977, **20**:364-366.
- 447 14. MacQueen J: **Some methods for classification and analysis of multivariate**
448 **observations.** *Proceedings of the fifth Berkeley symposium on mathematical statistics*
449 *and probability* 1967, **1**:281-297.
- 450 15. Lee DD, Seung HS: **Learning the parts of objects by non-negative matrix**
451 **factorization.** *Nature* 1999, **401**:788-791.

- 452 16. De Sousa E Melo F, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LPMH, de Jong JH, de
453 Boer OJ, van Leersum R, Bijlsma MF, et al: **Poor-prognosis colon cancer is defined by**
454 **a molecularly distinct subtype and develops from serrated precursor lesions.**
455 *Nature Medicine* 2013, **19**:614-618.
- 456 17. Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, Di Narzo AF, Yan P,
457 Graeme Hodgson J, Weinrich S, et al: **Gene expression patterns unveil a new level of**
458 **molecular heterogeneity in colorectal cancer.** *Journal of Pathology* 2013, **231**:63-
459 76.
- 460 18. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC,
461 Schiappa R, Guenot D, Ayadi M, et al: **Gene expression classification of colon cancer**
462 **into molecular subtypes: characterization, validation, and prognostic value.** *PLoS*
463 *Medicine* 2013, **10**.
- 464 19. Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, Snel MH, Chresta
465 CM, Rosenberg R, Nitsche U, et al: **Colorectal cancer intrinsic subtypes predict**
466 **chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal**
467 **transition.** *International Journal of Cancer* 2013, **134**:552-562.
- 468 20. Schlicker A, Beran G, Chresta CM, McWalter G, Pritchard A, Weston S, Runswick S,
469 Davenport S, Heathcote K, Castro DA, et al: **Subtypes of primary colorectal tumors**
470 **correlate with response to targeted treatment in colorectal cell lines.** *BMC*
471 *Medical Genomics* 2012, **5**:1-15.
- 472 21. Rousseeuw PJ: **Silhouettes: A graphical aid to the interpretation and validation of**
473 **cluster analysis.** *Journal of Computational and Applied Mathematics* 1987, **20**:53-65.
- 474 22. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve
475 RM, Qian Z, Ryder T, et al: **Genomic and transcriptional aberrations linked to**
476 **breast cancer pathophysiologies.** *Cancer Cell* 2006, **10**:529-541.
- 477 23. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich
478 R, Geisler S, et al: **Repeated observation of breast tumor subtypes in independent**
479 **gene expression data sets.** *Proc Natl Acad Sci USA* 2003, **100**:8418-8423.
- 480 24. Raghavan UN, Albert R, Kumara S: **Near linear time algorithm to detect community**
481 **structures in large-scale networks.** *Physical Review E* 2007, **76**:36106.
- 482 25. Onken MD, Worley LA, Ehlers JP, Harbour JW: **Gene Expression Profiling in Uveal**
483 **Melanoma Reveals Two Molecular Classes and Predicts Metastatic Death**
484 **Advances in Brief Gene Expression Profiling in Uveal Melanoma Reveals Two**
485 **Molecular Classes and Predicts Metastatic Death.** *Cancer Res* 2004:7205-7209.
- 486 26. Onken MD, Worley LA, Dávila RM, Char DH, Harbour JW: **Prognostic testing in uveal**
487 **melanoma by transcriptomic profiling of fine needle biopsy specimens.** *The*
488 *Journal of molecular diagnostics : JMD* 2006, **8**:567-573.
- 489 27. Worley LA, Onken MD, Person E, Robirds D, Branson J, Char DH, Perry A, Harbour JW:
490 **Transcriptomic versus chromosomal prognostic markers and clinical outcome in**
491 **uveal melanoma.** *Clinical Cancer Research* 2007, **13**:1466-1471.
- 492 28. Laurent C, Valet F, Planque N, Silveri L, Maacha S, Anezo O, Hupe P, Plancher C, Reyes C,
493 Albaud B, et al: **High PTP4A3 phosphatase expression correlates with metastatic**
494 **risk in uveal melanoma patients.** *Cancer Research* 2011, **71**:666-674.
- 495 29. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by**
496 **shrunk centroids of gene expression.** *Proceedings of the National Academy of*
497 *Sciences* 2002, **99**:6567-6572.
- 498 30. Triozzi PL, Achberger S, Aldrich W, Crabb JW, Saunthararajah Y, Singh AD: **Association**
499 **of tumor and plasma microRNA expression with tumor monosomy-3 in patients**
500 **with uveal melanoma.** *Clinical Epigenetics* 2016, **8**:80.

- 501 31. Sokal RR, Rohlf FJ: **The Comparison of Dendrograms by Objective Methods.** *Taxon*
502 1962, **11**:33-40.
- 503 32. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A,
504 Pomeroy SL, Golub TR, Lander ES, others: **Gene set enrichment analysis: a**
505 **knowledge-based approach for interpreting genome-wide expression profiles.**
506 *Proceedings of the National Academy of Sciences* 2005, **102**:15545-15550.
- 507 33. Wilkerson MD, Hayes DN: **ConsensusClusterPlus: a class discovery tool with**
508 **confidence assessments and item tracking.** *Bioinformatics* 2010, **26**:1572-1573.
- 509 34. Gaujoux R, Seoighe C: **A flexible R package for nonnegative matrix factorization.**
510 *BMC Bioinformatics* 2010, **11**:367.
- 511 35. Csárdi G, Nepusz T: **The igraph software package for complex network research.**
512 *InterJournal Complex Systems* 2006, **1695**:1-9.
- 513 36. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K: **cluster: Cluster Analysis**
514 **Basics and Extensions.** R package version 2.0.4. edition; 2016.
- 515 37. Therneau T: **A package for survival analysis in S.,** R package version 2.37-4 edition;
516 2014.
- 517

518 **Figure Legends**

519 **Figure 1. An overview of our pipeline for cluster reconciliation.** Gene expression – or other
520 equivalently structured molecular data – is input as a genes by samples matrix. This data is then fed
521 through multiple consensus clustering algorithms (in this case, HC, KM and NMF) to produce
522 multiple clustering solutions. These are then reconciled to create “subtype communities” of similar
523 clusters from across the algorithms’ solutions, by applying community detection to networks
524 representing the similarity between clusters from all the algorithms.

525

526 **Figure 2. Breast cancer subtypes and their association with intrinsic subtypes – application of**
527 **polyCluster. (A-B)** Similarity of each set of clusters generated by consensus A) NMF, B) HC and
528 C) KM to the known breast cancer subtypes of each sample (as assigned by correlation to PAM
529 centroids) using 118 breast cancer samples from a published dataset [22]. A hypergeometric test
530 was used to test the significance of overlap between the clusters and the known subtypes. bNMF,
531 bHC and bKM represent NMF, HC and KM subtypes, respectively. Norm – normal-like, lumA –
532 luminal A and lumB – luminal B subtypes.

533

534 **Figure 3. Subtype communities of breast cancer identified using polyCluster. (A)** A
535 hypergeometric (HYP) test and **(B)** PMI was used to assess the significance of the overlap between
536 each pair of clusters using Chin breast cancer data set. The resulting FDR corrected p values were
537 plotted as edge colours/weights in this network, with each node representing a cluster. The size of
538 each node represents the number of samples that cluster contains, and those nodes in a lighter shade
539 represent clusters with associations to known subtypes that are not significant (FDR corrected $p >$
540 0.05). Gray shading marks dense groups of clusters as defined by network community detection.
541 bHYP and bPMI represent HYP and PMI subtype breast cancer communities, respectively.

542

543 **Figure 4. Subtype communities of uveal melanoma identified using polyCluster.** (A) A
544 hypergeometric (HYP) test and (B) PMI was used to assess the significance of the overlap between
545 each pair of clusters using uveal melanoma data set. The resulting FDR corrected p values were
546 plotted as edge colours/weights in this network, with each node representing a cluster. The size of
547 each node represents the number of samples that cluster contains, and those nodes in a lighter shade
548 represent clusters with associations to known subtypes that are not significant (FDR corrected $p >$
549 0.05). Gray shading marks dense groups of clusters as defined by network community detection.
550 mHYP and mPMI represent HYP and PMI subtype melanoma communities, respectively.

551

552

553 **Figure 5. GSEA enrichment plots** of (A) the mHYP1 uveal melanoma community, showing
554 significant enrichment of immunity-related gene sets, and (B) the mHYP3 uveal melanoma
555 community, showing significant enrichment of neural-related gene sets.

556

557 **Figure 6. Prognosis and GSEA analysis of uveal melanoma subtype communities.** (A-B)
558 Metastasis-free survival in the A) discovery and B) validation cohorts, respectively, was significantly
559 different between communities. (C-D) GSEA enrichment plots of C) the mHYP1 and (B) the HYP3
560 uveal melanoma communities, showing significant enrichment of class 2 published subtypes.,

561

Figure 1

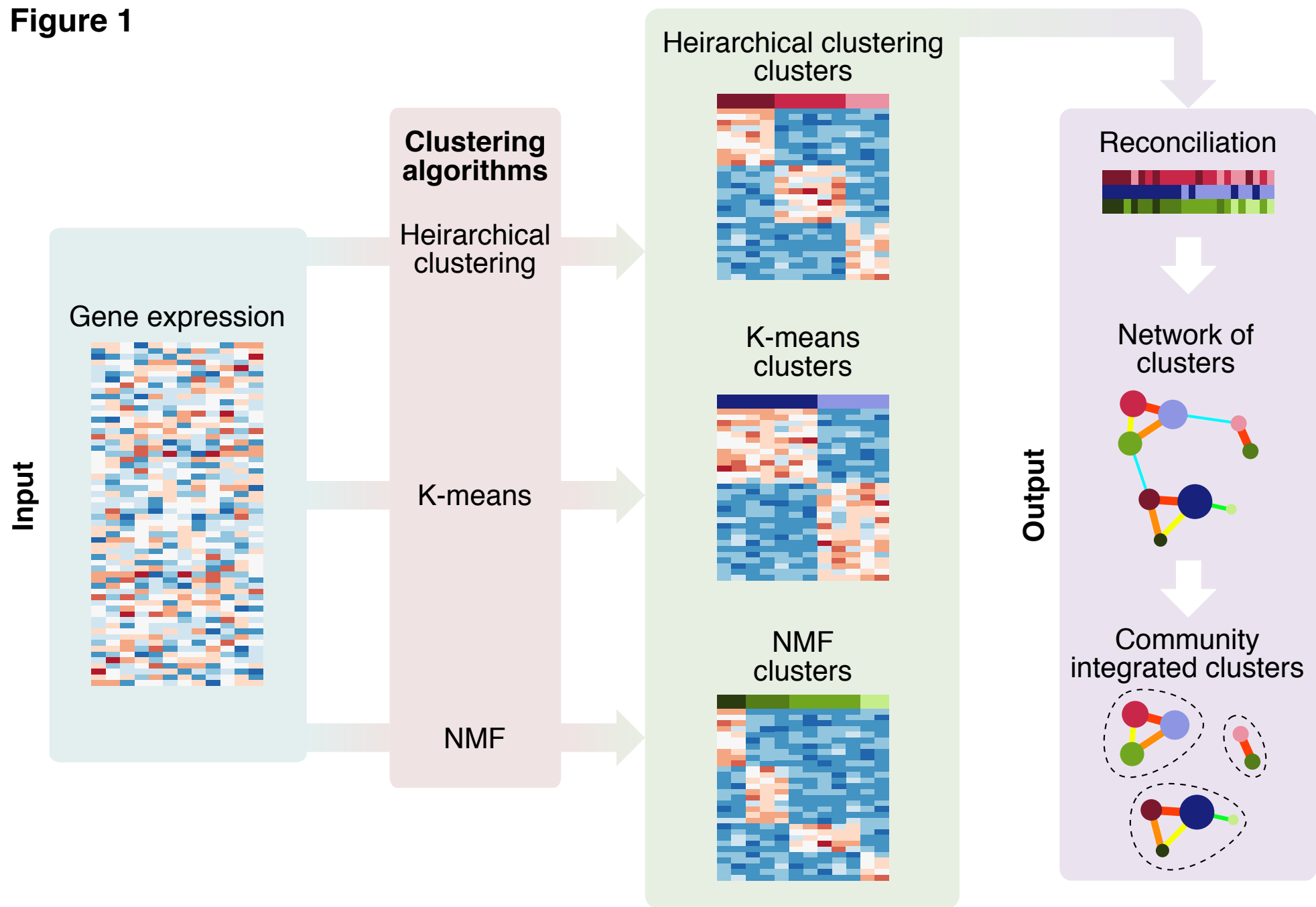


Figure 2

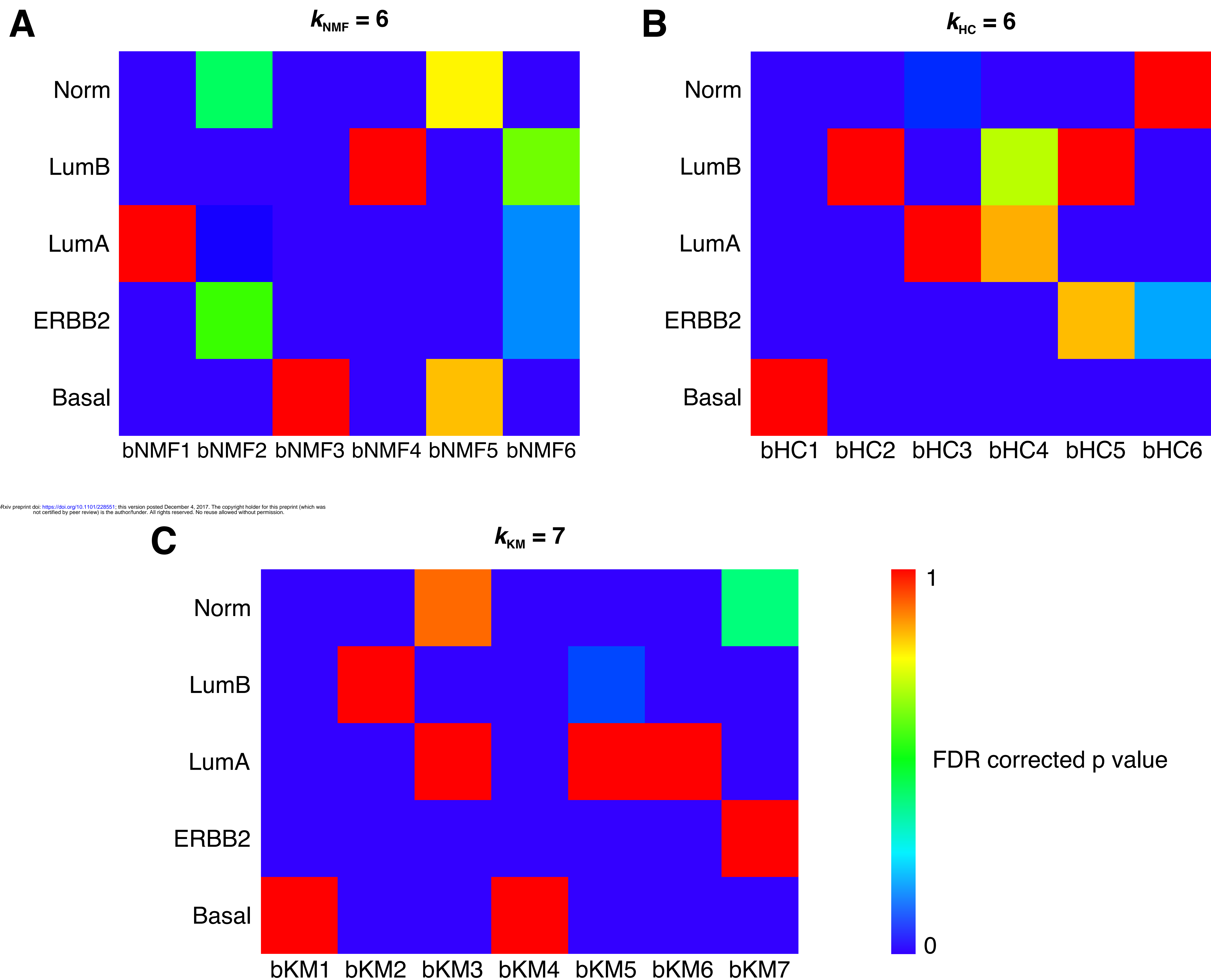
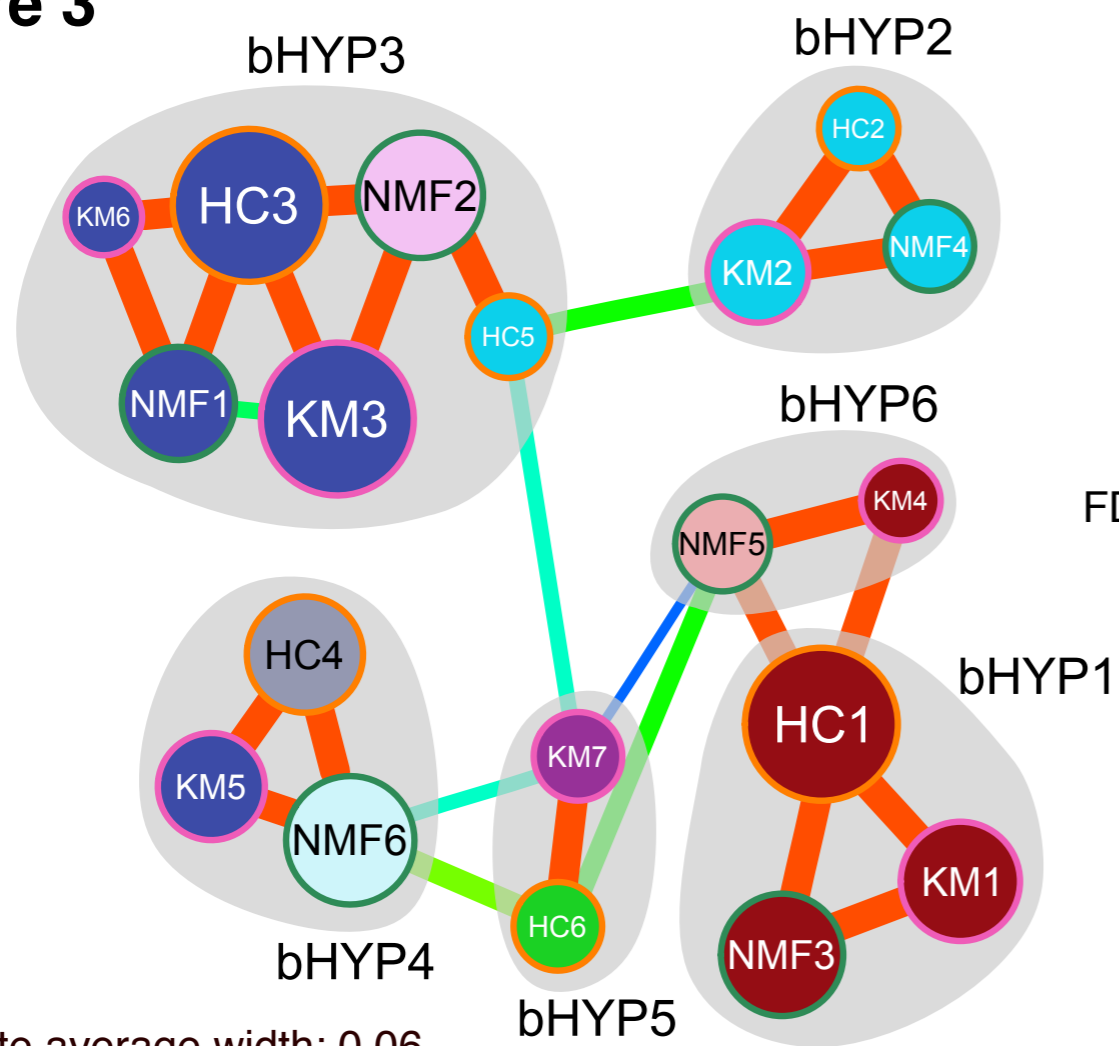


Figure 3**A**

FDR corrected p value

 ≤ 0.05 > 0.05

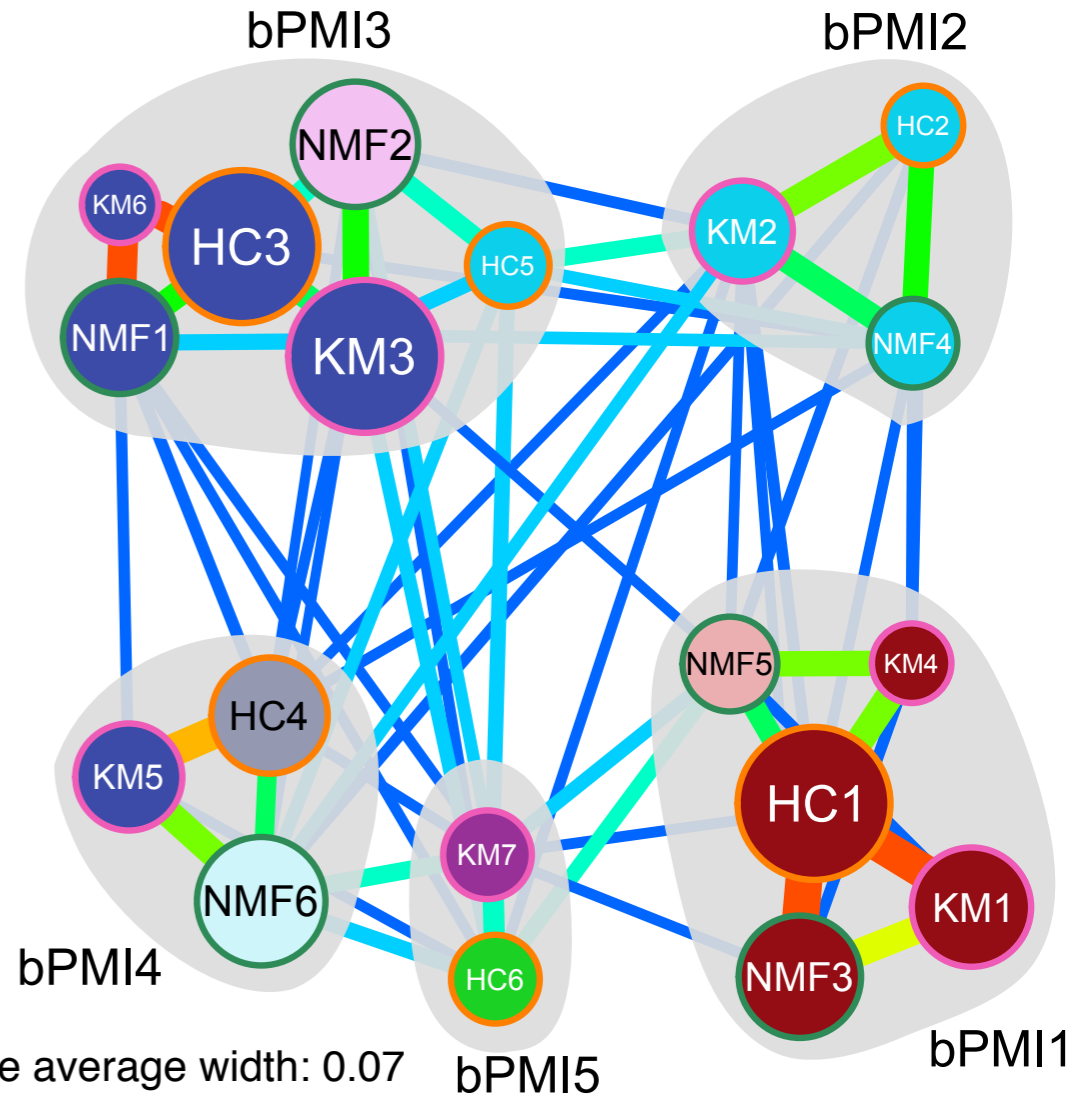
- Basal
- ERBB2
- Lum A
- Lum B
- Norm

- HC
- KM
- NMF

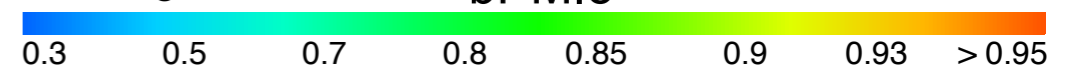
Silhouette average width: 0.06



FDR Corrected P-Value

B

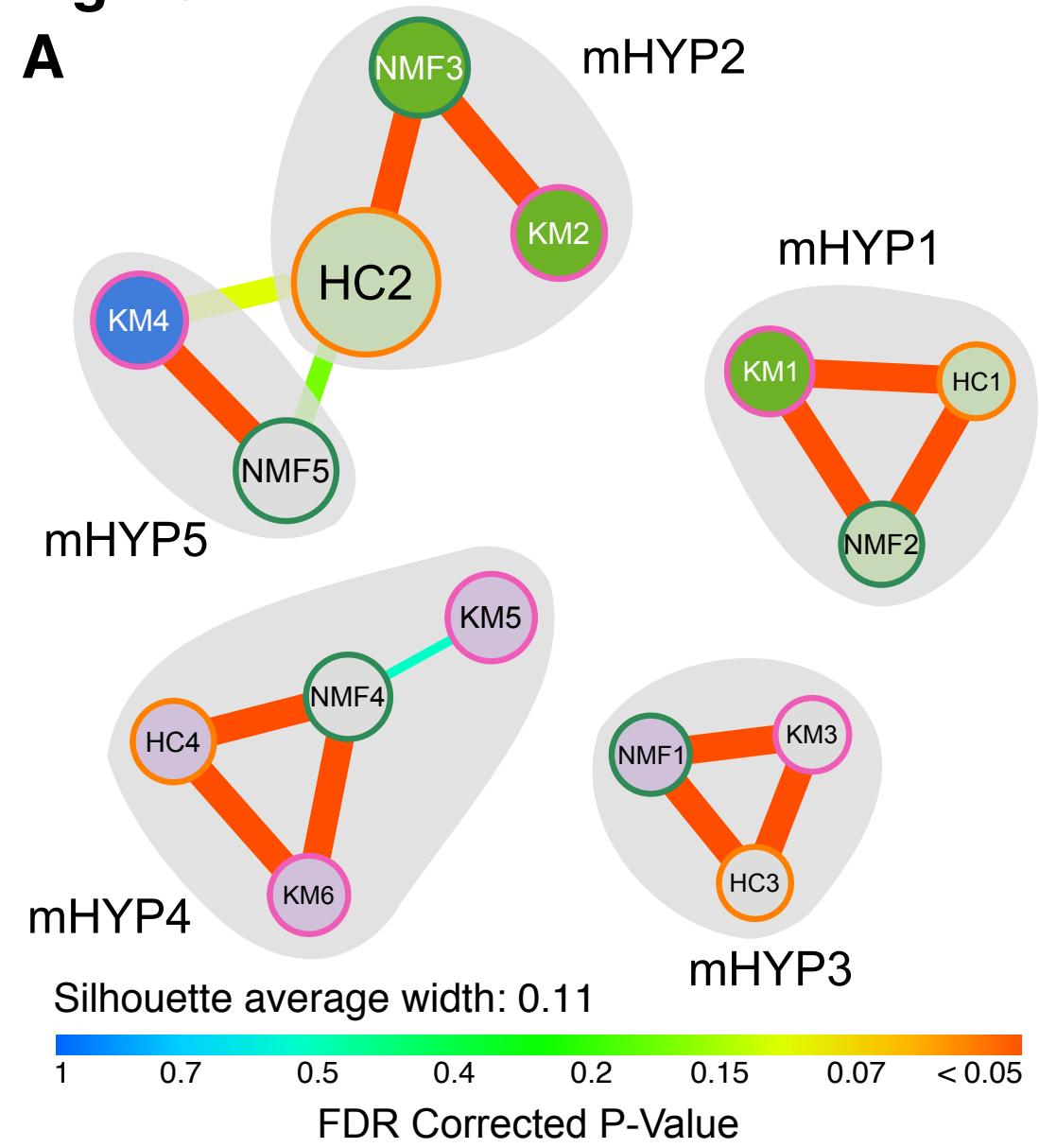
Silhouette average width: 0.07



Proportion of Maximum Intersection (PMI)

Figure 4

A



B

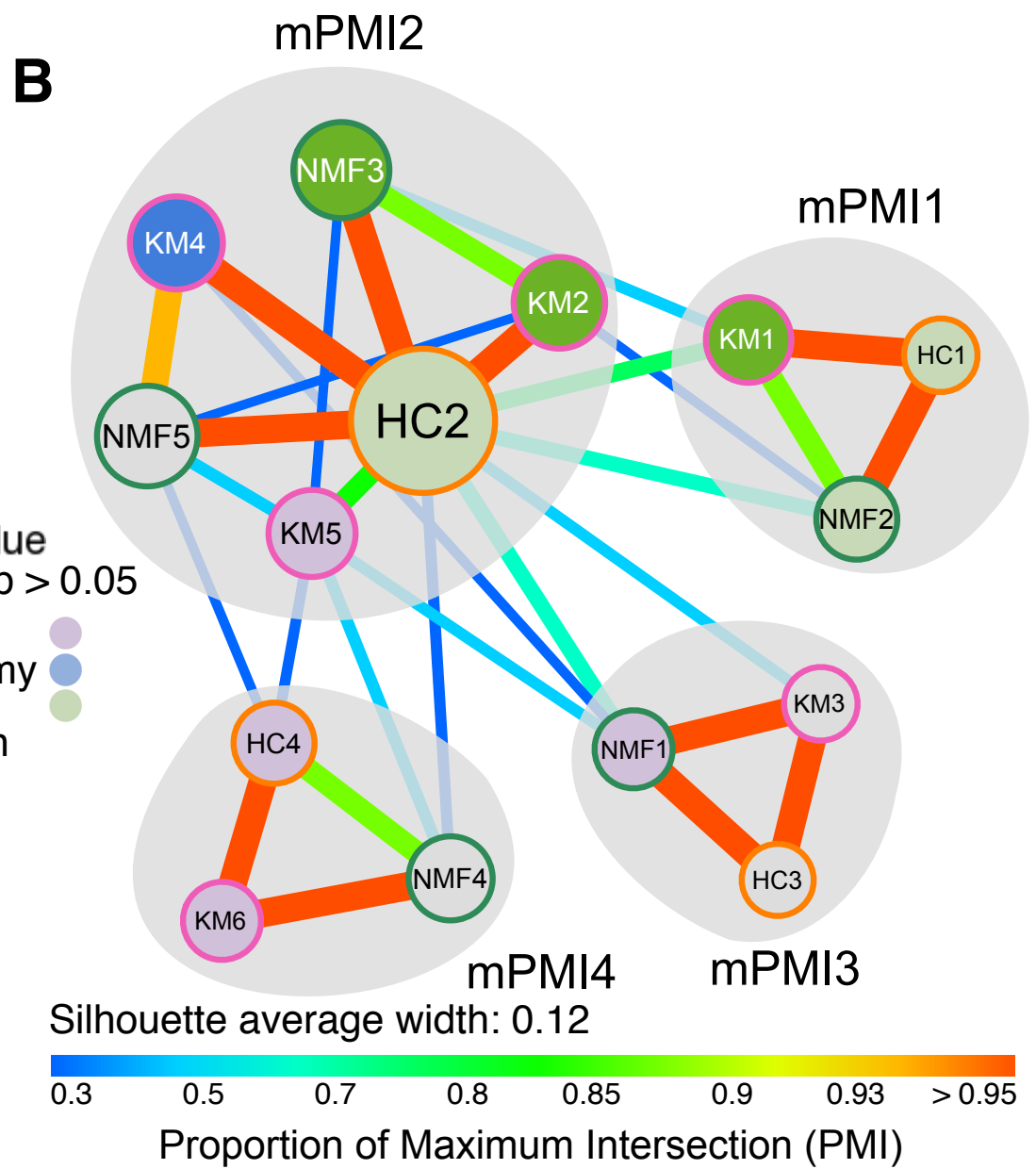
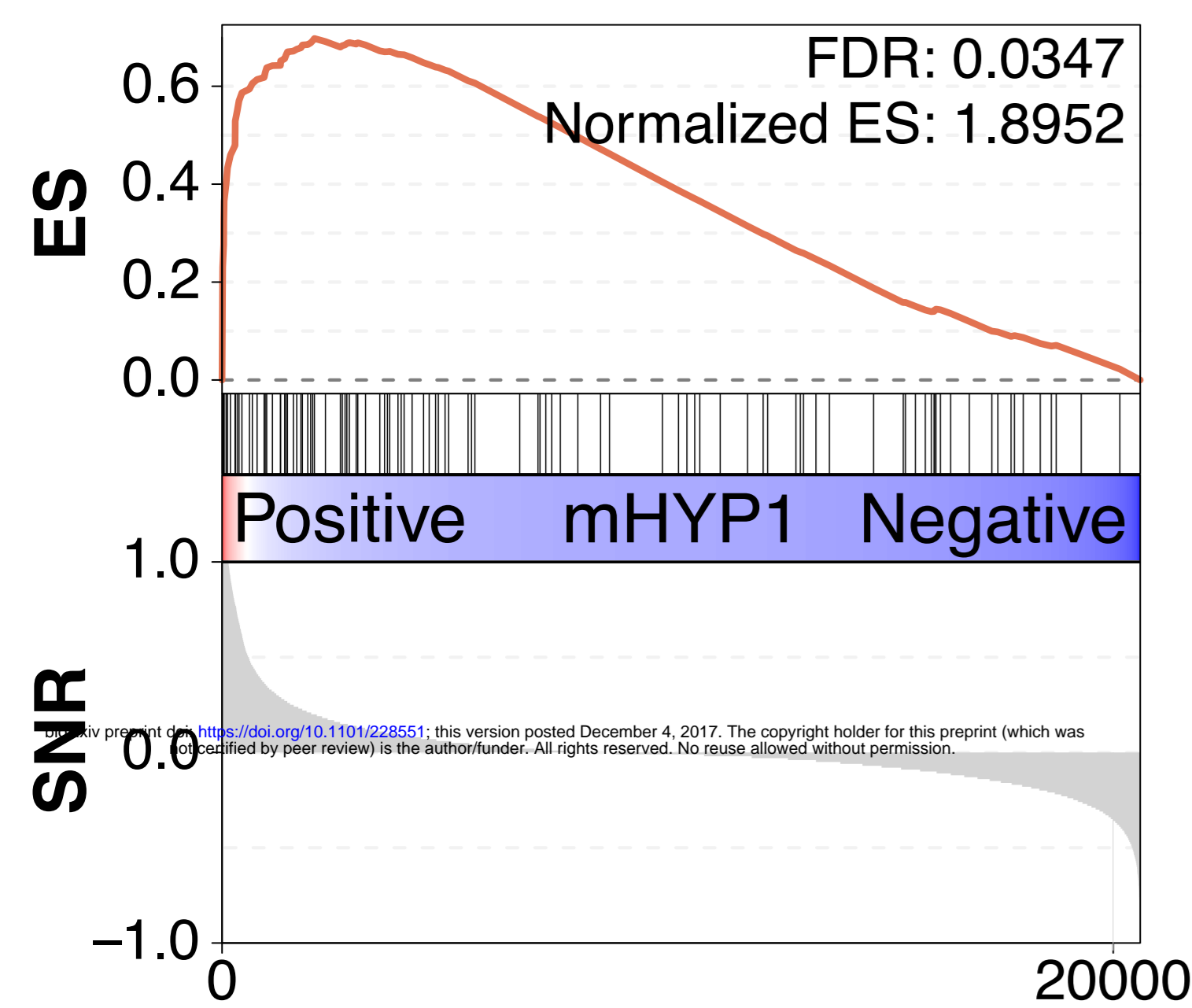


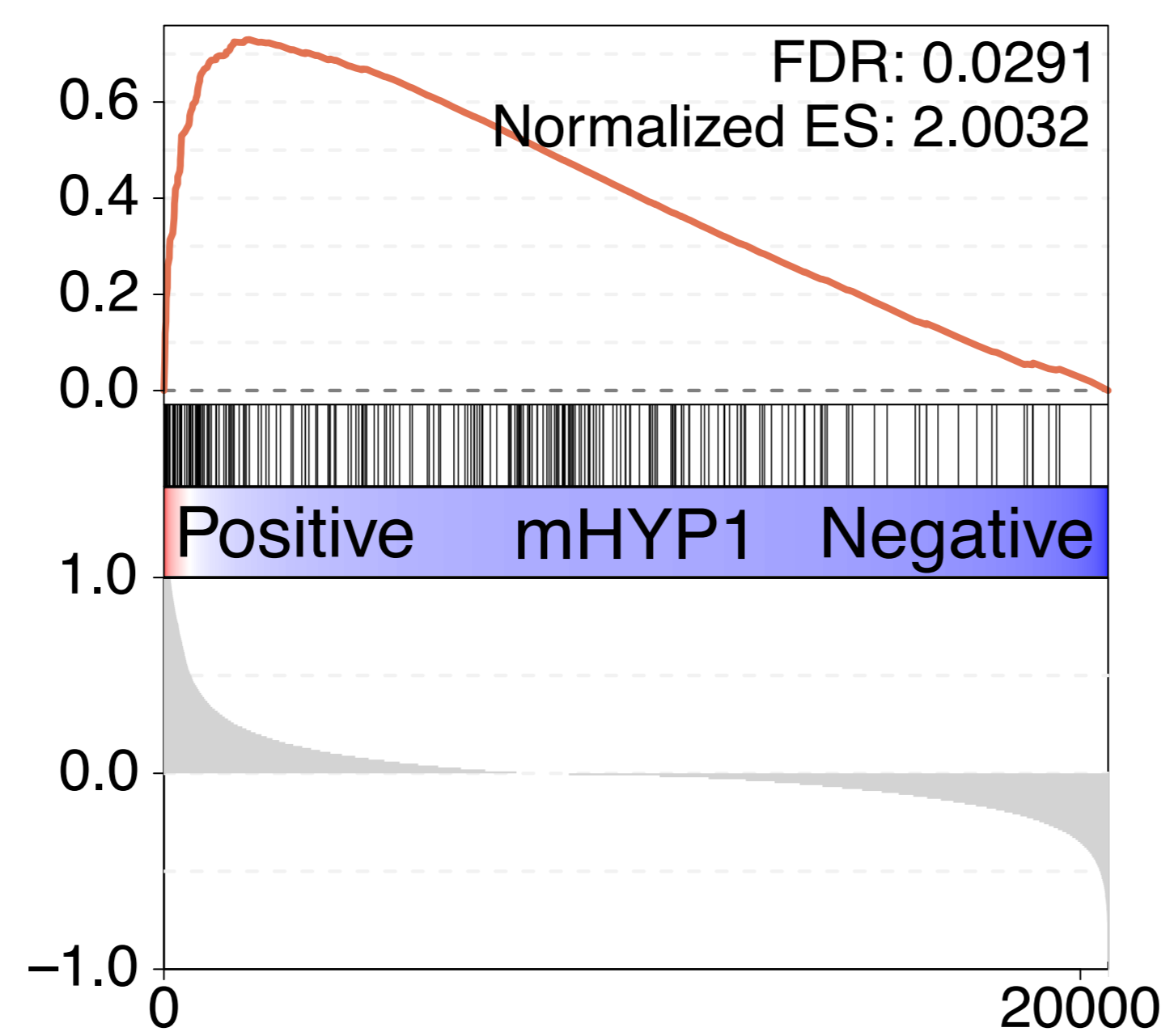
Figure 5

A

KEGG: T-Cell Receptor Signaling Pathway

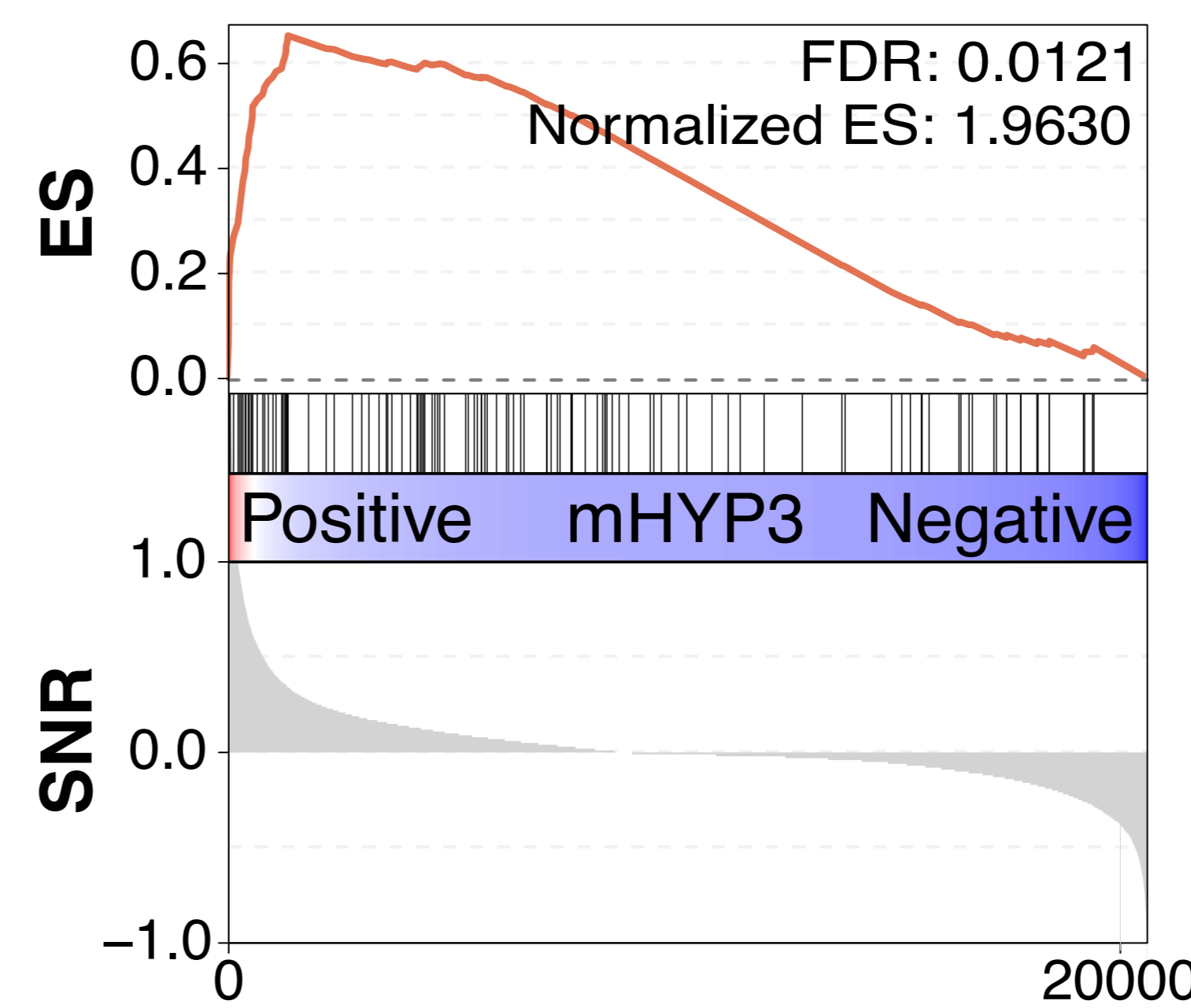


KEGG: Cytokine Cytokine Receptor Interaction

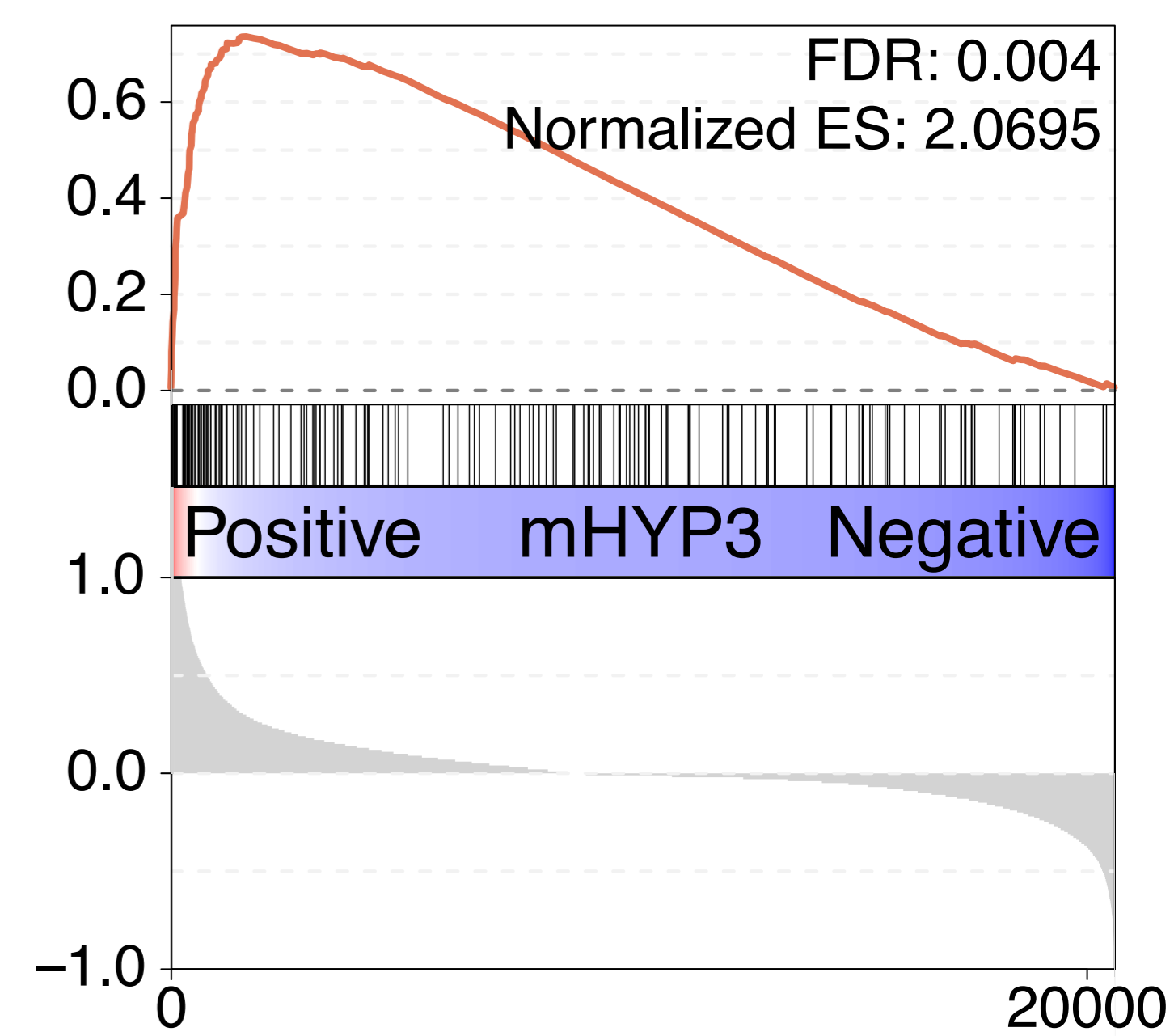


B

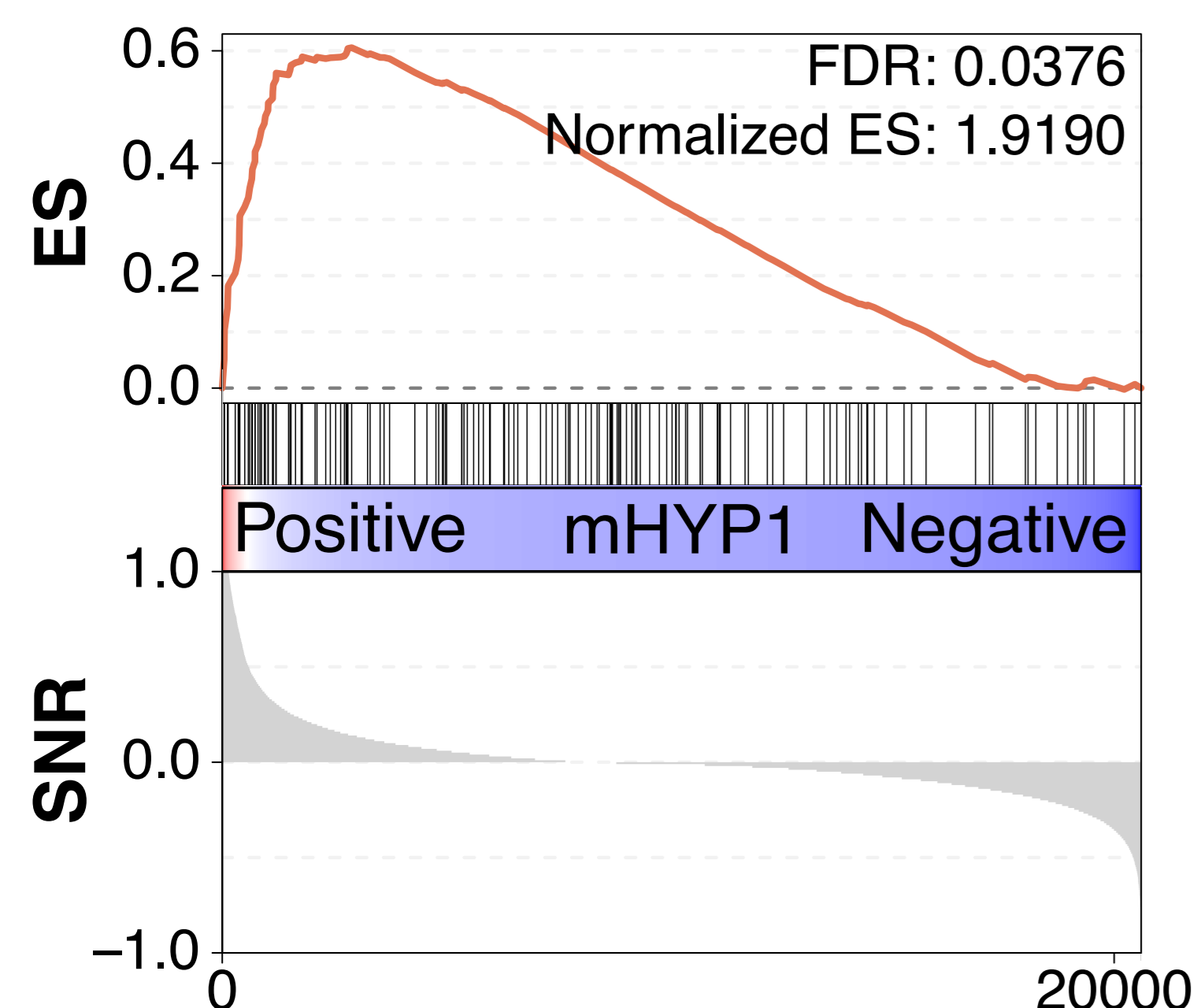
Verhaak: Glioblastoma Neural



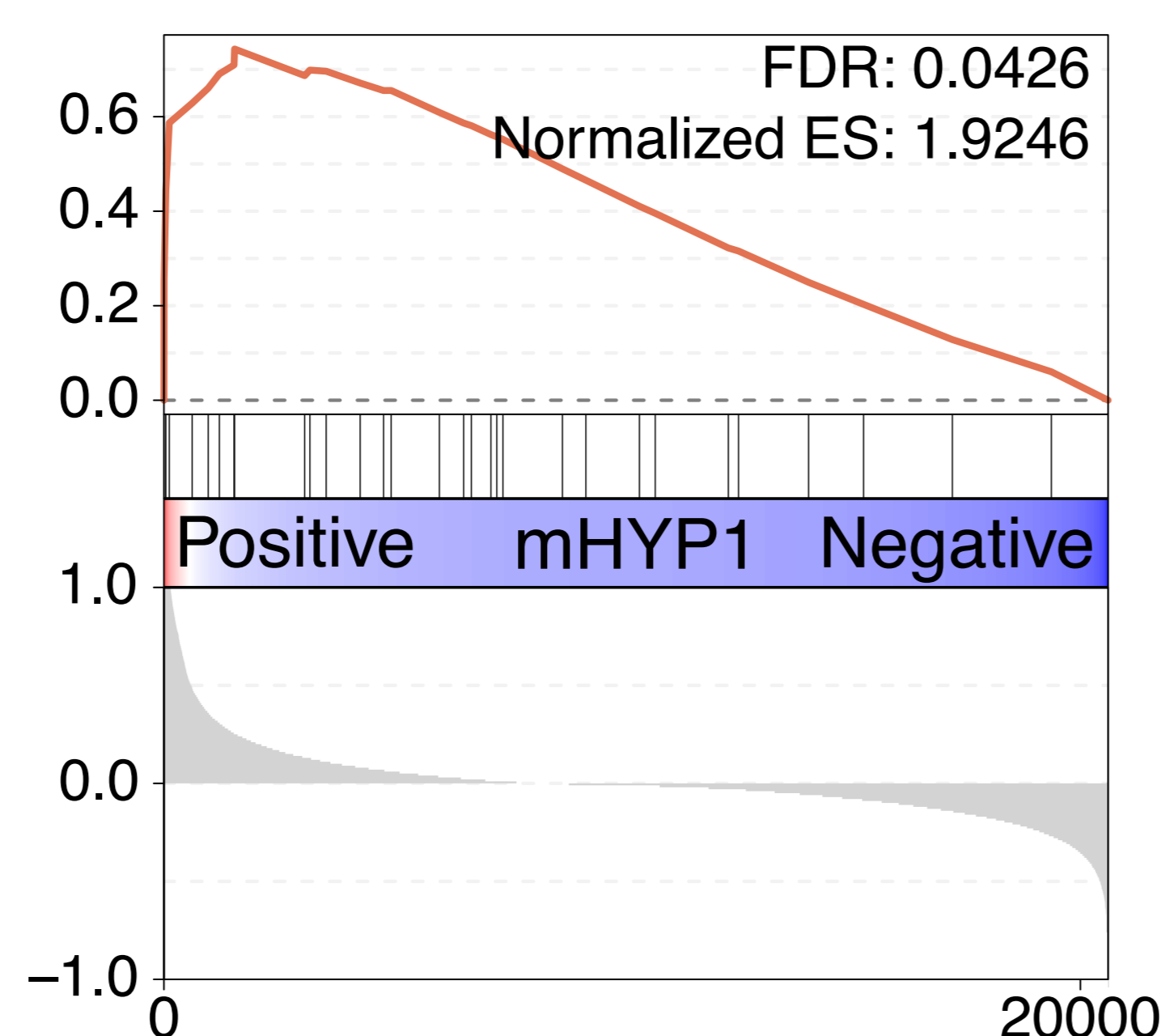
Reactome: Transmission Across Chemical Synapses



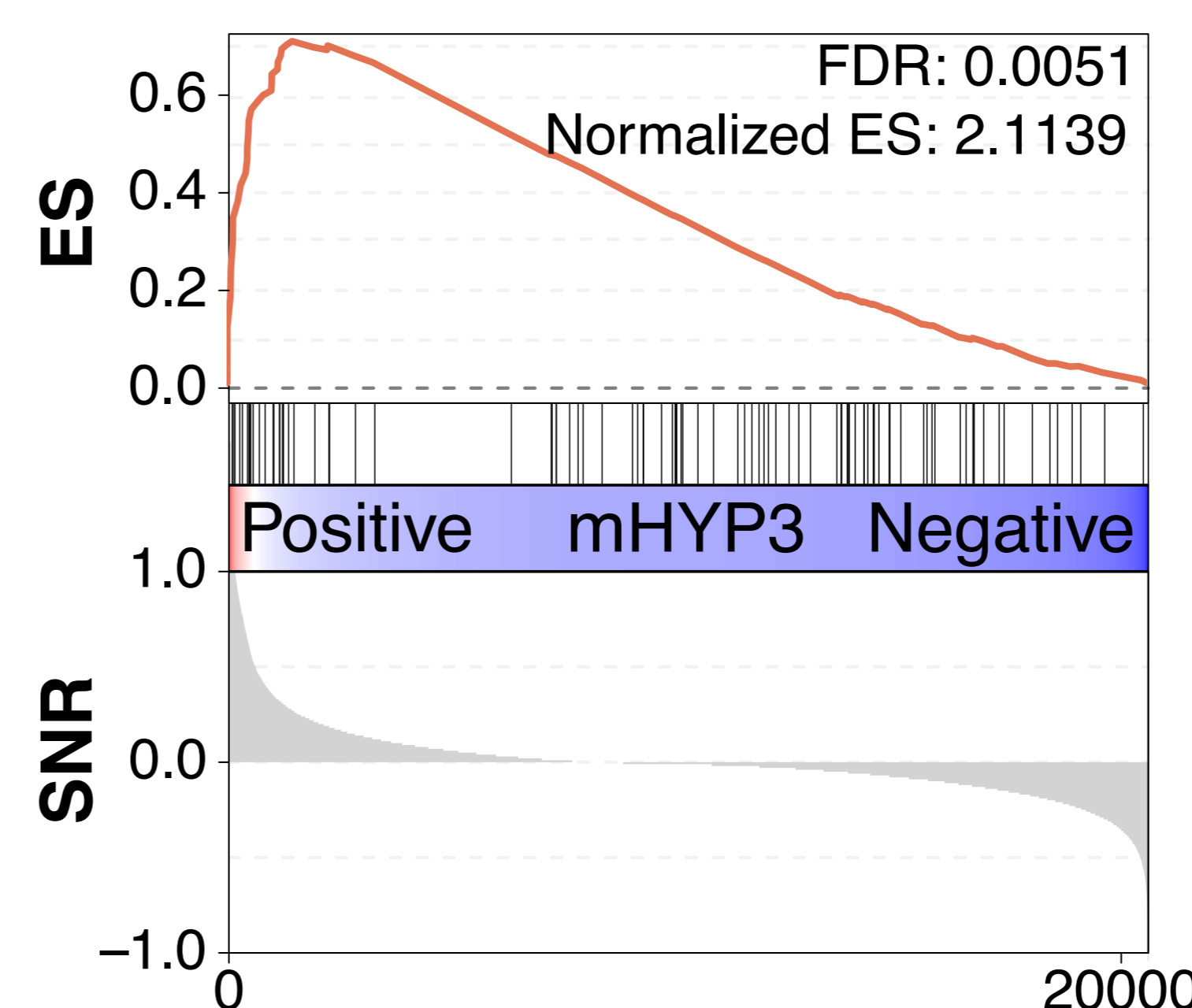
KEGG: JAK STAT Signaling Pathway



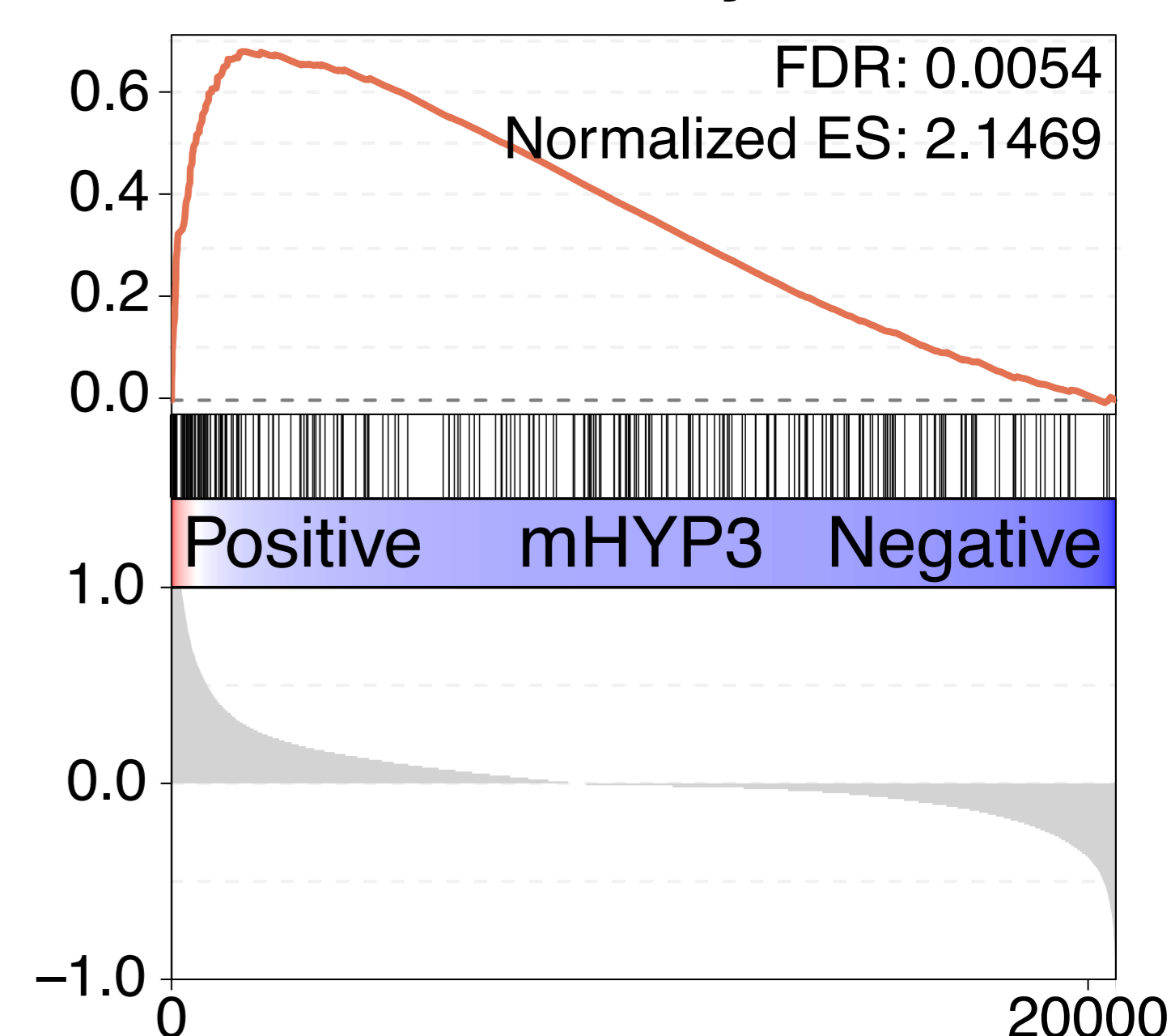
PID: IL2 STAT5 Pathway



Reactome: Potassium Channels



Reactome: Neuonal System



Rank in ordered gene list

Rank in ordered gene list

Rank in ordered gene list

Rank in ordered gene list

Figure 6

