Supplementary material for "Multi-project and Multi-profile joint Non-negative Matrix Factorization for cancer omic datasets"

D.A. Salazar, N. Pržulj and C.F. Valencia

July 31, 2021

1 Supplementary Sections

S1 Datasets Pre-processing

The omic profiles used were gene expression, miRNA expression, gene-level Copy Number variation, and sensitivity profile (1). We paired the projects to obtain the same set of variables (genes, miRNAs, and genes in CNV). The miRNA (Reads Per Million) and gene (normalized RSEM) expression profiles were in the pseudo logarithmic transformation for both projects. The values in the CNV profile in the CCLE project correspond to the average copy number of the genomic region of a gene. Therefore, these values are log2 transformed with a pseudo count of 1. Similarly, in the TCGA project, the CNV profile obtained by TCGA-Assembler also reports gene-level copy number value, but without transforming, we calculated the log2 transformed with a pseudo count of 1.

Table 1: Common dimensions for TCGA and CCLE projects. For the CCLE project, only the drug sensitivity profile was used. Additionally, the platforms used for each omic profile are included.

	Profiles			
	Gene expression	miRNA expression	Copy Number Variation	Area Under the Curve Drug sensitivity profile
Number of features in common	15227	314	8754	262
TCGA platforms	RNAseq Illumina HiSeq	HiSeq.hg19.mirbase20 Illumina HiSeq	cnv.hg19 Affymetrix SNP Array 6.0	-
CCLE platforms	Illumina HiSeq 2000 or HiSeq 2500	Nanostring platform	Illumina HiSeq 2000 or Illumina GAIIX	Replicate collapsed logfold change values relative to DMSO

We performed some corrections on the data to eliminate features, reduce the effect of outliers and impute data:

1. We calculated the asymmetry by variables in each profile. First, we trimmed the variables with asymmetry above 2, i.e., we replaced column values above the 99^{th} percentile with that percentile value. While if the value was below -2, all values below the 1^{st} percentile were replaced with that value. In this way, we removed the outliers.

2. We calculated the missing values using the softImpute v1.4.0 package, an iterative method for matrix completion that uses nuclear-norm regularization. In our case, we used the iterative soft-thresholded svds to impute the missing values.

3. We removed features that had a variance of zero or close to zero. We used the Scikit-learn v0.24.2 package, which has the VarianceThreshold function to remove variables based on a threshold. For the mRNA and CNV profiles, we used a threshold of 0.0099 since there are genes with low variability. However, since the other profiles have small dimensions compared to the mRNA or CNV profiles, we decided not to apply this filter.

Therefore, the matrices between projects had the same features and within each project had the same objects (patients or cell lines). To ensure a positive input profiles, we scaled the values per

columns using the formula $x_{ij}-X_{min}/X_{max}-X_{min}$ where x_{ij} is the i^{th} observation in the j^{th} column of the matrix X. X_{max} and X_{min} are the maximum and the minimum value of the j^{th} column, respectively.

S2 Constraints description

- mRNA-mRNA interaction constraints ($\Theta^{(t)}_{gen}$). For the first gene constraint (Θ^{1}_{gene}), genetic interactions of genes were identified using the BioGRID v3.5 database (Oughtred et al. 2019). The second gene constraint (Θ^{2}_{gene}) was obtained from STRINGdb v9.1 package (Franceschini et al. 2012), which refers to the association or physical interaction between genetic products. The third gene constraint (Θ^{3}_{gene}) was obtained from KEGG DB using graphite v1.32.0 package (Sales et al. 2012), which refer to the interaction between enzymes in metabolic pathways. The last gene constraint (Θ^{4}_{gene}) was generated using limma v3.42.2 (Ritchie et al. 2015). This last constraint was generated as follows. First, we identified differential gene expression by log-fold changes; then, we generated the network of differentially regulated genes between normal tissue and primary tumor tissue. From the TCGA-Assembler tool, we download normal tissue from patients with glioblastoma in TCGA project.

- miRNA-miRNA interaction constraints (Θ_{miRNA}). These constraints were obtained from CancerNet and represent the synergism that exists between miRNAs to regulate various processes such as tumorigenesis, angiogenesis, and metastasis (Meng et al. 2015).

- miRNA-Target interaction constraints ($R_{miRNA-target}$). These constraints were obtained from the miRNet database (Fan and Xia 2018). The $R_{mRNA-miRNA}$ constraint represents regulation in gene expression, i.e., the mRNA-miRNA interaction that occurs in altered processes in cancer. In addition, miRNA-drug constraint ($R_{miRNA-drug}$) was obtained from miRNet v2.0 (Chang et al. 2020). These constraints contain information of miRNAs that down-regulate genes that are targets of drugs.

- Drug-drug interaction constraints (Θ_{drug}). These constraints were obtained from the Drugbank v5.0 (Wishart et al. 2018) which contains drug interaction information, i.e., concomitant use of drugs generates an adverse event.

S3 Co-cluster assignment rules

Co-clusters from H_I

We decided to create a rule of assignment of molecules in the clusters that also contemplates the maximum values belonging to a specific cluster, but that discourages the molecule from repeatedly appearing in several clusters, as conventional methods assign a maximum value to each cluster based on a threshold. In general, the method searches the maximum value by rows in the matrix H and chooses them if their membership is high (more significant than the second quartile). In addition, this process is repeated for the second maximum value using the third quartile.

From the H_I matrices, where the rows are the k clusters and the columns are the m_I molecules, e.g., genes. First, We identified where each molecule had its maximum value. For example, if $k = 90$ and the maximum value for the $gene_{32}$ was in the row 84. Then, this gene was incorporated into the cluster 84. Once we identified the vector of genes that belonged to a specific cluster, we only took into account the genes with a high value of belonging; this is, only genes that were above the second quartile (Q_2) belong to the k cluster. On the other hand, we repeated the same procedure for the second maximum values of the H_I matrix, but its incorporation was more rigorous and was only allowed for genes that were above the Q_3. We repeated this process for the different omic profiles, which established the co-clusters, i.e., the belonging of a group of genes, miRNA, and drugs to a specific cluster (Supplementary Figure F5).

Patient cluster membership

As the W matrix dimensions contain the patients or cell lines (rows) and the clusters (columns), we can incorporate a set of patients into a cluster, which we assigned depending on their belonging value (given by Q_3). However, we identified clusters with few patients, and we reassigned them to clusters with more than 30 samples; this reassignment was made considering the proximity of their value to that of the compared cluster.

For TCGA, we could use information such as the progression-free interval (J. Liu et al. 2018) to contrast if there are differences in the probability of survival. The latter was verified because PFI is a better measure for checking the aggressiveness of low-grade glioma. We constructed the Kapan-Meier survival plots and used the log-rank test to confirm that the populations come from different distributions. In addition, we compared the M&M-jNMF patient clusters with predefined clinical groups the histological and molecular classification for low-grade glioma used in this study was performed by (Ceccarelli et al. 2016; Brat et al. 2015).

S4 Description of synthetic data

We created two artificial datasets G_1 and G_2 with three profiles in common. The samples for both datasets were 150, and the three profiles contained 750, 1400, and 1250 variables. We set the original range (k) at 5. The pairwise associations between observations belonging to a cluster were defined as follows: firstly, in each column of the matrix W, a value of 10 was randomly assigned, i.e., the observations with this value belong to cluster k. Each observation belongs to only one cluster. For example, for a $k = 10$, we assigned randomly the value of 10 to some observations to the k^{th} column in W matrix ($s \times k$). Secondly, we filled the matrix W with values obtained from a uniform distribution $[0, 1]$. Similarly, we performed the same procedure for each row in the H_I matrix. In both cases, the associations correspond to the pairs of individuals or variables in a cluster; for example, there will be ten associations if there are five observations in a cluster.

S5 Comparison patient groups obtained using M&M-jNMF

We identified important differences between the groups obtained by our method. In Tables 2, 3, and 4, we present the genes that differed between the groups of patients. We compared the group of patients whose survival curves were very close, so we present genes or markers that can make this difference between the different groups. Then, we compared between groups: I and III, III and IV, and V and VII.

Table 2: Top 6 of differentially expressed genes between group I and group III TCGA patients. The cBioportal tool was used to identify these genes.

Gene	Cytoband	Mean in group I	Mean in group III	Standard deviation in group I	Standard deviation in group III	Log Ratio	q-Value	Higher expression in
LHX5	12q24.13	5.56	3.39	2.40	2.22	2.17	2.25e-10	Group I
TLX1	10q24.31	6.35	4.42	2.44	2.65	1.93	4.64e-07	Group I
TRIM67	1q42.2	10.33	8.47	2.40	1.95	1.85	2.96e-09	Group I
TESPA1	12q13.2	3.51	8.84	2.40	1.88	-5.32	1.74e-44	Group III
WIF1	12q14.3	3.06	8.18	2.42	1.72	-5.11	7.48e-46	Group III
KCNS1	20q13.12	3.96	8.95	2.36	1.46	-4.99	3.88e-51	Group III

Table 3: Top 6 of differentially expressed genes between group III and group IV TCGA patients. The cBioportal tool was used to identify these genes.

Gene	Cytoband	Mean in group III	Mean in group IV	Standard deviation in group III	Standard deviation in group IV	Log Ratio	q-Value	Higher expression in
TESPA1	12q13.2	8.84	2.37	1.88	1.91	6.47	3.63e-30	Group III
KCNS1	20q13.12	8.95	2.67	1.46	2.31	6.28	1.34e-23	Group III
KCNV1	8q23.2	8.01	2.09	1.05	2.05	5.92	3.89e-24	Group III
LEFTY2	1q42.12	3.68	7.04	2.21	1.76	-3.36	8.49e-15	Group IV
TNFSF12-TNFSF13	17p13.1	2.80	6.01	2.13	2.46	-3.21	5.01e-10	Group IV
FCGBP	19q13.2	8.66	11.81	1.66	1.81	-3.15	1.62e-14	Group IV

Table 4: Top 6 of differentially expressed genes between group VII and group V TCGA patients. The cBioportal tool was used to identify these genes.

Gene	Cytoband	Mean in group VII	Mean in group V	Standard deviation in group VI	Standard deviation in group V	Log Ratio	q-Value	Higher expression in
KLRC2	12p13.2	8.33	4.30	2.23	2.07	4.02	4.42e-12	Group V
PAX1	20p11.22	4.41	0.89	2.54	1.47	3.52	2.17e-10	Group V
HMX1	4p16.1	6.82	3.35	1.98	1.82	3.48	9.83e-12	Group V
TCTEX1D1	1p31.3	2.59	8.06	1.46	2.42	-5.47	8.62e-17	Group VII
MOXD1	6q23.2	4.93	10.32	2.67	1.72	-5.39	8.35e-16	Group VII
LGR6	1q32.1	3.55	8.83	2.58	2.43	-5.28	6.64e-14	Group VII

Among these groups, some traits can differentiate disease progression. For example, among group III, there is a high expression of the WIF1 gene. This gene is related to the suppression of senescence in glioblastoma (Lambiv et al. 2011). While in group I, we found a high expression of the TRIM67 gene. This gene has been related to the promotion of cell apoptosis through the

NF-κB pathway in lung cancer cells (R. Liu et al. 2019). We explored the over-expressed genes in group III because there are more than 640 genes enriched in integral components of the plasma membrane, glycoproteins, and membrane proteins, i.e., this group corresponds to a type of tumor cells with a higher transmembrane activity than those of group I.

Among group III and group IV, we found 79 highly expressed genes in group IV. These genes are involved in angiogenesis and immune system processes; for example, the TNFSF12 gene has a negative relationship with overall survival in glioma patients, as it promotes invasive properties of glial cells (Portela, Mitchell, and Casas-Tintó 2020).

Finally, group V and group VII also show differences in gene expression. Group VII has 378 genes with high expression with processes related to cell movement and components of the extracellular space. Cluster VII has a clear difference since it is much more active with the extracellular environment. A relevant gene in this group is the LGR6 gene which has been linked to chemoresistance in ovarian cancer (Ruan et al. 2019). Other highly expressed genes that promote drug excretion are also highly expressed in this cluster, for example, ALDH1A3, ALDH3A1, GSTM5, and MAOB.

S6 Comparison between cell lines groups and patient groups

We identified 7 groups of patients (I-VII) and 9 groups of cell lines (1-9). We measured the PCASimilarity score to determine similarities in signaling pathways, metabolic pathways, and gene ontology (biological process, molecular function, and cellular component) terms. Accordingly, we compared 121 terms between cell line groups and patient groups (Supplementary Figure F6); we removed some terms not in line with glioma or cancer, e.g., infectious diseases. From these terms and the PCASimilarity score measured, we found that the proportion in which these terms clustered between-group 5 cell lines and patient groups was 29% (test for equality of proportions $p - value < 0.001$), i.e., the proportions are not equal in at least two groups (Table 5).

Table 5: Proportions between patient groups and cell line groups for the 107 terms (GO and KEGG). N corresponds to the total number of cell line group vs. patient group pairs with a high PCASimilarity score for any enriched term. For example, there are 185 pairs between-group 5 and all patient groups.

Group of cell lines	Overall, N = 647	Group 1, N = 60 (9.3%)	Group 2, N = 44 (6.8%)	Group 3, N = 51 (7.9%)	Group 4, N = 39 (6.0%)	Group 5, N = 185 (29%)	Group 6, N = 57 (8.8%)	Group 7, N = 86 (13%)	Group 8, N = 75 (12%)	Group 9, N = 51 (7.9%)
PCASimilarity score	0.95 (0.04)	0.94 (0.04)	0.94 (0.05)	0.95 (0.04)	0.94 (0.04)	0.96 (0.04)	0.96 (0.04)	0.96 (0.04)	0.94 (0.04)	0.95 (0.05)
Groups of patients										
I	101 (16%)	9 (15%)	7 (16%)	7 (14%)	5 (13%)	32 (17%)	4 (7.0%)	12 (14%)	17 (23%)	8 (16%)
II	93 (14%)	7 (12%)	7 (16%)	3 (5.9%)	4 (10%)	33 (18%)	8 (14%)	13 (15%)	14 (19%)	4 (7.8%)
III	97 (15%)	8 (13%)	8 (18%)	6 (12%)	4 (10%)	26 (14%)	13 (23%)	16 (19%)	8 (11%)	8 (16%)
IV	88 (14%)	9 (15%)	3 (6.8%)	11 (22%)	6 (15%)	22 (12%)	8 (14%)	12 (14%)	9 (12%)	8 (16%)
V	88 (14%)	10 (17%)	8 (18%)	10 (20%)	8 (21%)	16 (8.6%)	11 (19%)	11 (13%)	8 (11%)	6 (12%)
VI	91 (14%)	9 (15%)	5 (11%)	5 (9.8%)	8 (21%)	28 (15%)	6 (11%)	11 (13%)	9 (12%)	10 (20%)
VII	89 (14%)	8 (13%)	6 (14%)	9 (18%)	4 (10%)	28 (15%)	7 (12%)	10 (12%)	10 (13%)	7 (14%)

[1] Mean (SD); n (%)

Similar metabolic pathways between-group 5 cell lines and patient groups corresponded to oxidative phosphorylation (I, II, III, IV, VI, and VII), mismatch repair (VI and VII), cell cycle (III, IV), DNA replication (II and III), DNA replication (II and III), Ribosome (II, III, VI, and VII), TNF signaling pathway (V), Wnt signaling pathway (II), apoptosis (VI), and Toll-like receptor signaling pathway (IV). Supplementary Figure F7 shows how expression can be similar/different between patient groups and cell lines. For example, Cytokine-cytokine receptor interaction has a low PCASimilarity score (0.66) between both projects, which was to be expected since immune system-related processes are primarily expressed in primary tumors by immune cell infiltration, a process that does not occur in cell lines (Yu et al. 2019). On the other hand, similar signaling pathways or metabolic pathways among almost all patient groups and group 5 corresponded to oxidative phosphorylation. Which in these cell lines may be activated by the culture medium, mainly Roswell Park Memorial Institute (RPMI) and Dulbecco's Modified Eagle Medium (DMEN), which increase the dependence of this pathway (Joly, Chew, and Graham 2021).

S7 Drug repurposing using drug sensitivity profile predicted for tumors

We found which genes, miRNAs, and metabolic pathways are associated with the sensible tumor groups. We identified that unique pathways whose genes have low expression correspond to oxidative phosphorylation, PPAR signaling pathway, carbon metabolism, and glucagon signaling pathway for all drugs. On the other hand, the pathways associated with genes up-regulated are phosphatidylinositol signaling system, sphingolipid signaling pathway, Wnt signaling pathway, TGF-beta signaling pathway, ErbB signaling pathway, FoxO signaling pathway, and JAK-STAT

signaling pathway. The case of oxidative phosphorylation is interesting since it is known that the activity of this pathway can increase drug resistance via autophagy. In a study by Lee et al. 2020 it was found that depletion of this pathway by inhibition of complex I and the ADLH enzyme reversed resistance to Irinotecan; as a consequence, the xenografts tumors decreased in size (Lee et al. 2020).

Genes and miRNAs also cluster differently between the sensitive and resistant groups. This situation means that among the drugs studied here, they have a different genetic and miRNA pattern. For example, the group of protein kinase inhibitors (-inib drugs) have molecular signatures different for each drug of this family; in some cases, some genes are highly expressed in resistant tumors. However, in other cases, the sensitive tumors show this pattern; for example, the TRIM67 gene is highly expressed in tumors sensitive to Pelitinib, and the opposite happens with the drug Erlotinib. This result is interesting given that this group of drugs, despite belonging to the protein kinase inhibitor group, their therapeutic targets can be very varied.

Finally, using mirNet v2.0 (Chang et al. 2020), we found that the miRNAs down-regulated in sensible tumors are involved in processes such as angiogenesis, Toll-like receptor signaling pathway, apoptosis, regulation of stem cell, and DNA damage response. For miRNAs up-regulated in sensible tumors, we found miRNAs the terms associated with DNA damage response, cell differentiation, response to hypoxia, and cell motility ($p-value < 0.05$). This group contains the miR-379 cluster which has 42 miRNAs (miR-889 family, miR-758 family, miR-668 family, miR-656 family, miR-543, miR-541 family, miR-539 family, miR-496, miR-495 family, miR-487 family, miR-410, miR-382 family, miR-381 family, miR-380 family, miR-379, miR-376a family, miR-329 family, miR-154 family, miR-134 family, miR-1197, and miR-1185-5p; $p-value < 0.05$) and miR-891a cluster which has 8 miRNAs (miR-892a, miR-892b, miR-891 family, miR-890, and miR-888 family; $p-value < 0.05$). The miRNA Clusters have gained importance in the study of diseases since they have fundamental roles in the development, regulation, metabolism, and others (Ghafouri-Fard et al. 2021). For instance, the miRNA miR-379/656 cluster is down-regulated in oligodendrogliomas. However, this cluster is known to have tumor suppressor roles in glioblastoma and glioneuronal tumors (Kumar et al. 2018). Therefore, we found that if there is an expression of this set of miRNAs, it is possible to increase the sensitivity to several drugs, including docetaxel, veliparib, trametinib, zibotentan, afatinib, linifanib, and vismodegib.

References

Brat, Daniel J. et al. (2015). "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas". In: *New England Journal of Medicine* 372.26, pp. 2481–2498. ISSN: 15334406. DOI: 10.1056/NEJMoa1402121.

Ceccarelli, Michele et al. (2016). "Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma". In: *Cell* 164.3, pp. 550–563. ISSN: 10974172. DOI: 10.1016/j.cell.2015.12.028.

Chang, Le et al. (July 2020). "miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology". In: *Nucleic Acids Research* 48.W1, W244–W251. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa467. URL: https://academic.oup.com/nar/article/48/W1/W244/5850315.

Fan, Yannan and Jianguo Xia (2018). "miRNet Functional Analysis and Visual Exploration of miRNA Target Interactions in a Network Context". In: pp. 215–233. DOI: 10.1007/978-1-4939-8618-7_10. URL: http://link.springer.com/10.1007/978-1-4939-8618-7%7B%5C_%7D10.

Franceschini, Andrea et al. (Nov. 2012). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration". In: *Nucleic Acids Research* 41.D1, pp. D808–D815. ISSN: 0305-1048. DOI: 10.1093/nar/gks1094. URL: http://www.ncbi.nlm.nih.gov/pubmed/23203871%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3531103%20http://academic.oup.com/nar/article/41/D1/D808/1057425/STRING-v91-proteinprotein-interaction-networks.

Ghafouri-Fard, Soudeh et al. (July 2021). "An update on the role of miR-379 in human disorders". In: *Biomedicine & Pharmacotherapy* 139, p. 111553. ISSN: 07533322. DOI: 10.1016/j.biopha.2021.111553. URL: https://linkinghub.elsevier.com/retrieve/pii/S0753332221003383.

Joly, James H., Brandon T. L. Chew, and Nicholas A. Graham (Apr. 2021). "The landscape of metabolic pathway dependencies in cancer cell lines". In: *PLOS Computational Biology* 17.4. Ed. by Jason W. Locasale, e1008942. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008942. URL: https://dx.plos.org/10.1371/journal.pcbi.1008942.

Kumar, Anupam et al. (Aug. 2018). "Identification of miR-379/miR-656 (C14MC) cluster downregulation and associated epigenetic and transcription regulatory mechanism in oligodendrogliomas". In: *Journal of Neuro-Oncology* 139.1, pp. 23–31. ISSN: 0167-594X. DOI: 10.1007/s11060-018-2840-6. URL: http://link.springer.com/10.1007/s11060-018-2840-6.

Lambiv, Wanyu L. et al. (July 2011). "The Wnt inhibitory factor 1 (WIF1) is targeted in glioblastoma and has a tumor suppressing function potentially by induction of senescence". In: *Neuro-Oncology* 13.7, pp. 736–747. ISSN: 1523-5866. DOI: 10.1093/neuonc/nor036. URL: https://academic.oup.com/neuro-oncology/article-lookup/doi/10.1093/neuonc/nor036.

Lee, Jae-Seon et al. (Sept. 2020). "Targeting Oxidative Phosphorylation Reverses Drug Resistance in Cancer Cells by Blocking Autophagy Recycling". In: *Cells* 9.9, p. 2013. ISSN: 2073-4409. DOI: 10.3390/cells9092013. URL: https://www.mdpi.com/2073-4409/9/9/2013.

Liu, Jianfang et al. (2018). "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics". In: *Cell* 173.2, 400–416.e11. ISSN: 10974172. DOI: 10.1016/j.cell.2018.02.052.

Liu, Rui et al. (July 2019). "TRIM67 promotes NF-κB pathway and cell apoptosis in GA-13315-treated lung cancer cells". In: *Molecular Medicine Reports*. ISSN: 1791-2997. DOI: 10.3892/mmr.2019.10509. URL: http://www.spandidos-publications.com/10.3892/mmr.2019.10509.

Meng, X et al. (Dec. 2015). "CancerNet: a database for decoding multilevel molecular interactions across diverse cancer types". In: *Oncogenesis* 4.12, e177–e177. ISSN: 2157-9024. DOI: 10.1038/oncsis.2015.40. URL: http://www.nature.com/articles/oncsis201540.

Oughtred, Rose et al. (2019). "The BioGRID interaction database: 2019 update". In: *Nucleic Acids Research* 47.D1, pp. D529–D541. ISSN: 0305-1048. DOI: 10.1093/nar/gky1079. URL: https://academic.oup.com/nar/article/47/D1/D529/5204333.

Portela, Marta, Teresa Mitchell, and Sergio Casas-Tintó (Jan. 2020). "Cell to cell communication mediates glioblastoma progression in Drosophila". In: *Biology Open*. ISSN: 2046-6390. DOI: 10.1242/bio.053405. URL: https://journals.biologists.com/bio/article/doi/10.1242/bio.053405/266457/Cell-to-cell-communication-mediates-glioblastoma.

Ritchie, Matthew E. et al. (Apr. 2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7, e47. ISSN: 1362-4962. DOI: 10.1093/nar/gkv007. URL: http://academic.oup.com/nar/article/43/7/e47/2414268/limma-powers-differential-expression-analyses-for.

Ruan, Xiaohong et al. (Sept. 2019). "Silencing LGR6 Attenuates Stemness and Chemoresistance via Inhibiting Wnt/β-Catenin Signaling in Ovarian Cancer". In: *Molecular Therapy - Oncolytics* 14, pp. 94–106. ISSN: 23727705. DOI: 10.1016/j.omto.2019.04.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S2372770519300506.

Sales, Gabriele et al. (2012). "graphite - a Bioconductor package to convert pathway topology to gene network". In: *BMC Bioinformatics* 13.1, p. 20. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-20. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-20.

Wishart, David S et al. (Jan. 2018). "DrugBank 5.0: a major update to the DrugBank database for 2018". In: *Nucleic Acids Research* 46.D1, pp. D1074–D1082. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1037. URL: http://academic.oup.com/nar/article/46/D1/D1074/4602867.

Yu, K. et al. (Dec. 2019). "Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types". In: *Nature Communications* 10.1, p. 3574. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11415-2. URL: http://www.nature.com/articles/s41467-019-11415-2.

2 Supplementary Figures

F1 Hyperparameters evaluation for the simulation case.

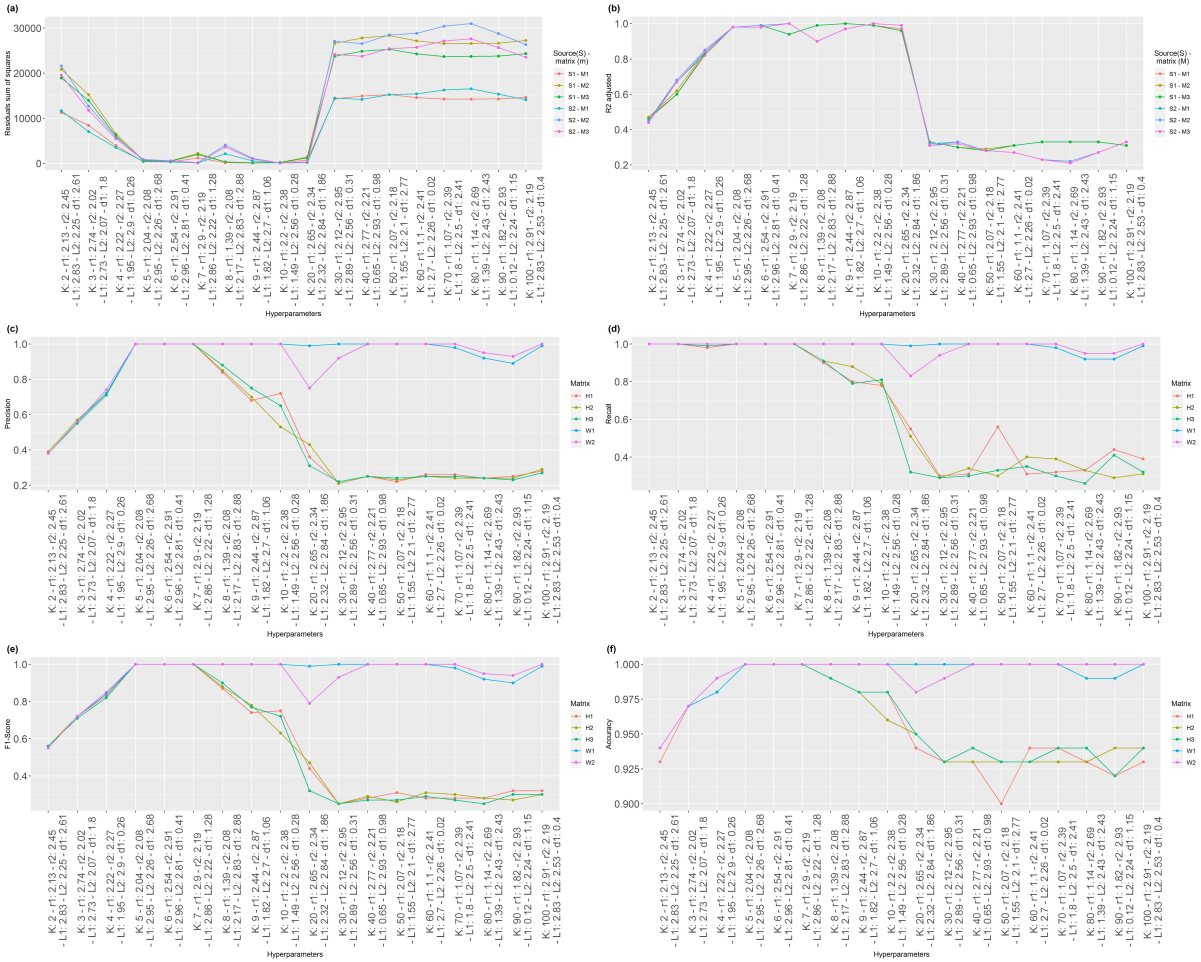


Figure 1: Tuning of hyperparameters to the M&MjNMF model for synthetic data. There are two sources (S), each with three input matrices (M). The factorization problem obtains five low-rank matrices (H and W). (a) Residual sum of squares between the predicted and original matrix; low values are expected, (b) $R^2_{adjusted}$ corresponds to the fit of the model data, (c) Precision, (d) Recall, (e) F1-Score, and (f) Accuracy.

F2 Metrics for choosing hyperparameters

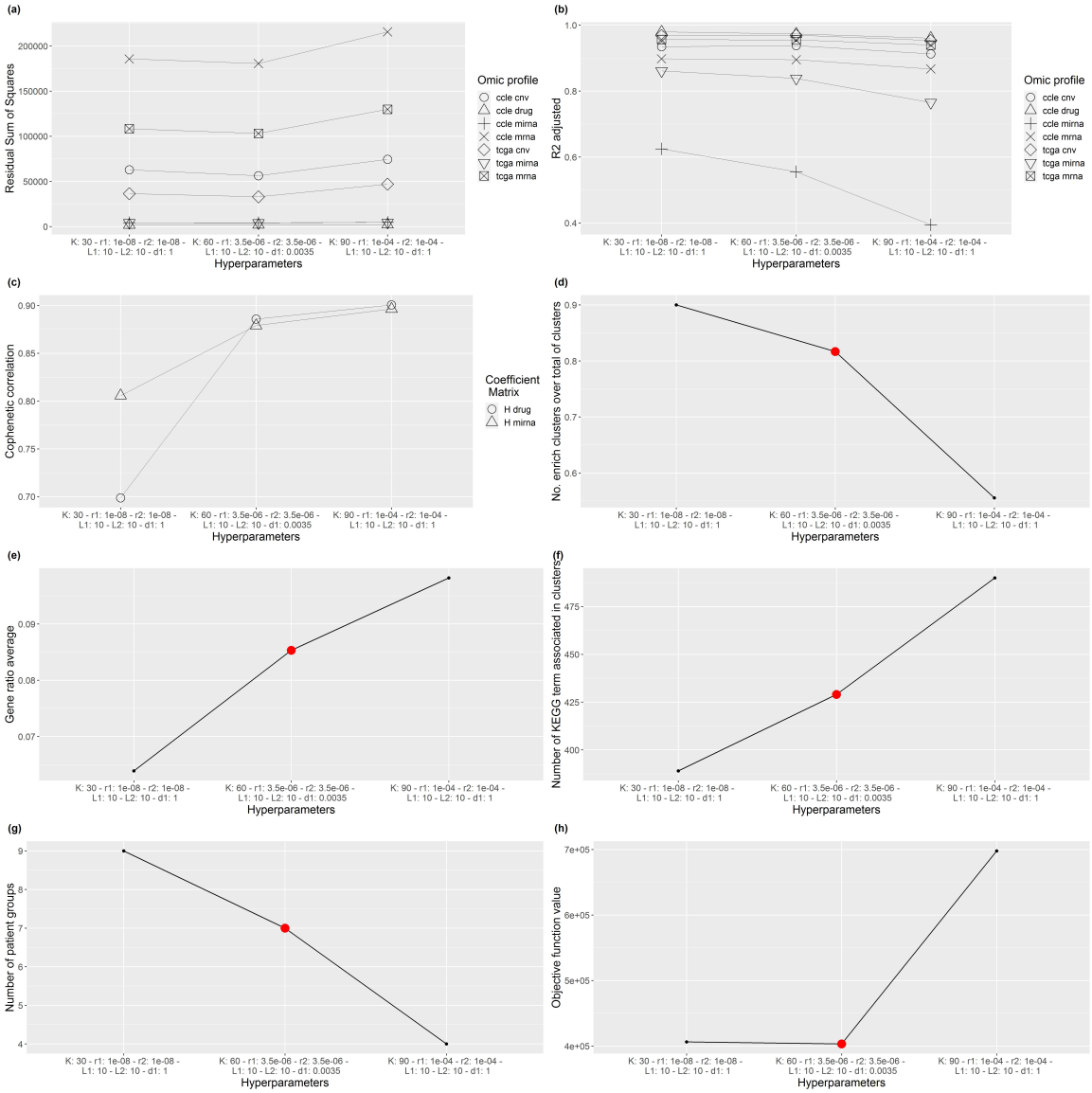


Figure 2: Metrics to choose an optimal set of hyperparameters for biological interpretation purposes. (a) Residual Sum of Squares, (b) $R^2_{adjusted}$, (c) Cophenetic coefficient for H_{drug} and H_{miRNA}, other profiles were computationally expensive, (d) the number of enriched clusters over total clusters (K) refers to the enriched clusters in KEGG terms, (e) the gene ratio corresponds to a proportion of how many genes belong to a given ontological term, (f) the number of KEGG terms associated with the enriched clusters, (g) the number of clusters for patients is expected to be greater than 3 because the standard clinical classification only includes three categories, and (h) the objective function value. The red dot is the set of Hyperparameters chosen.

F3 Gene enrichment dotplot

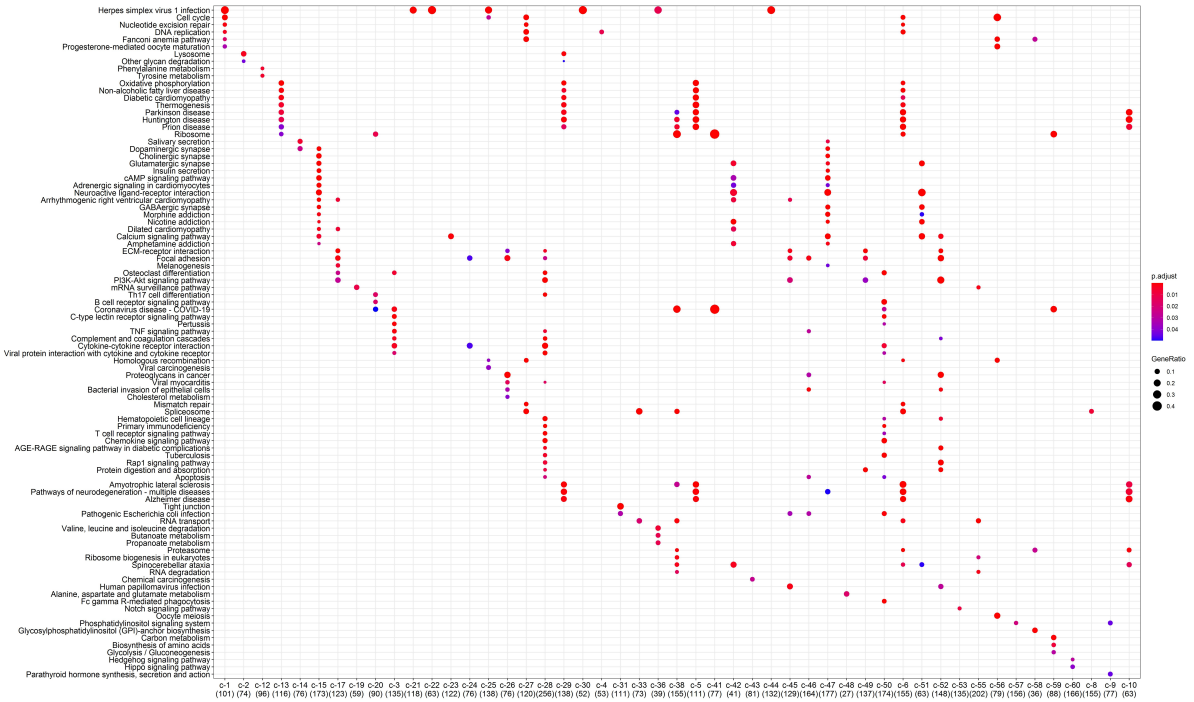


Figure 3: Dot plot of enrichment terms (X-axis) for gene clusters (Y-axis)

F4 Number of clusters similar between clusters of cell lines and clusters of patients.

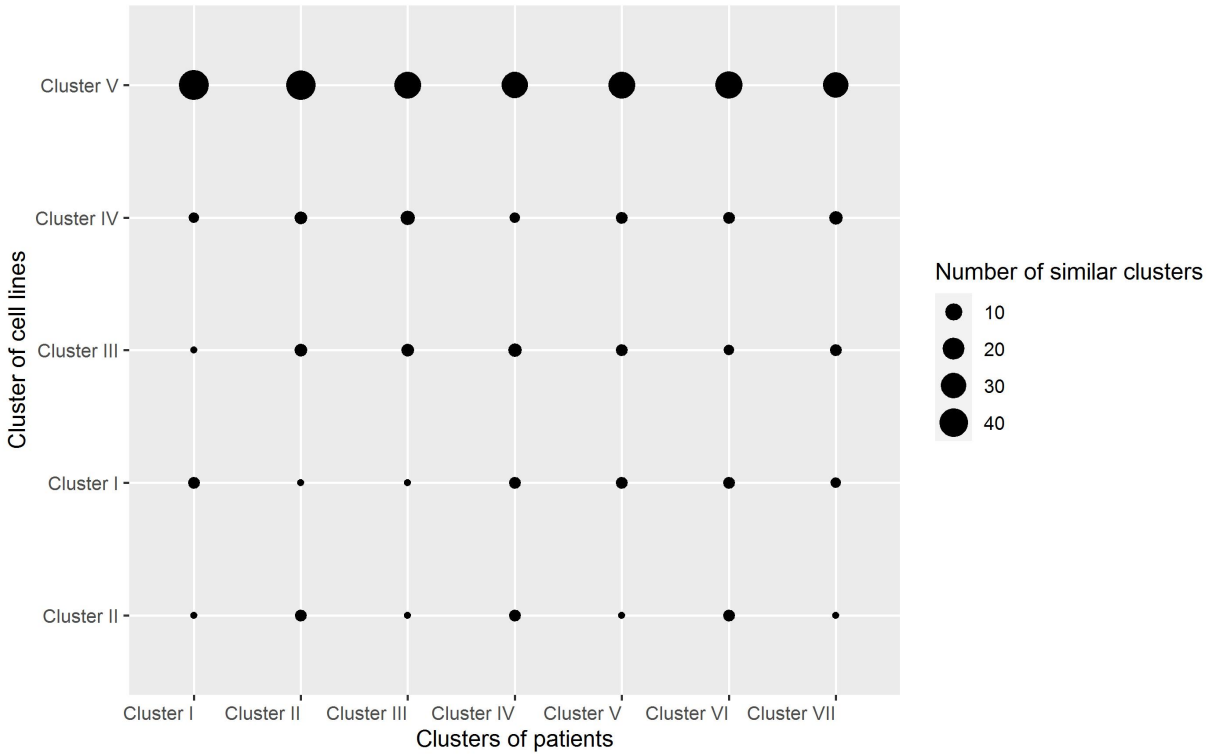


Figure 4: Matching between patient clusters and cell line clusters using PCASimilarity score. Similar mRNA, miRNA, and CNV clusters between cell lines and patients were identified.

F5 Molecule assignment rule scheme for each cluster.

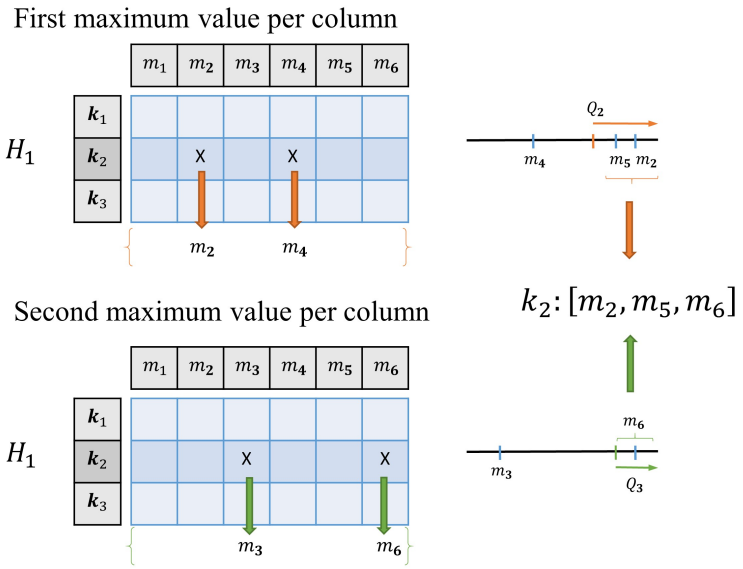
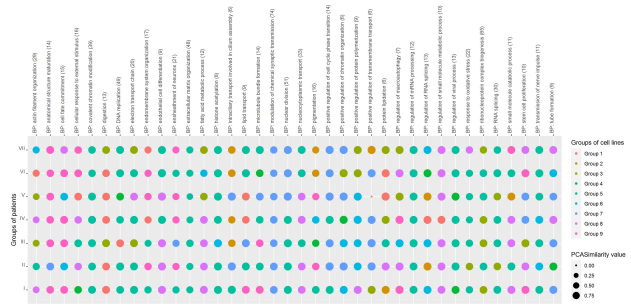
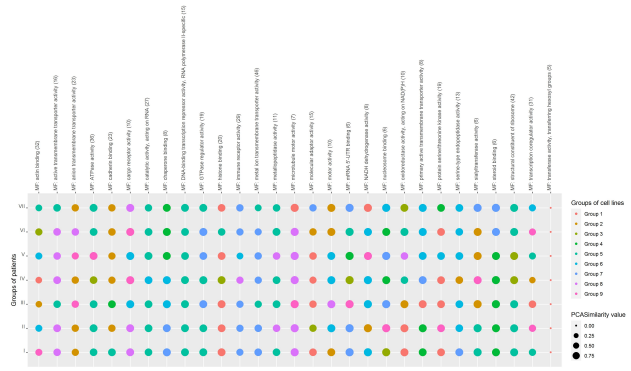


Figure 5: Co-cluster creation using H_I.

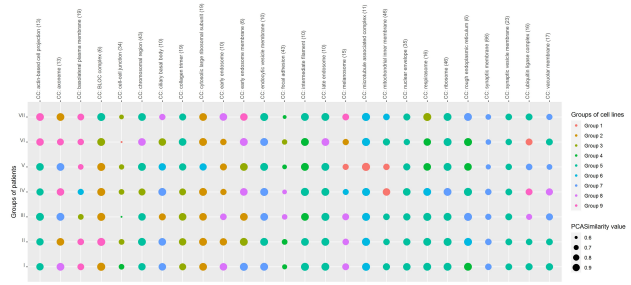
F6 Similarities between cell line groups and tumor groups.



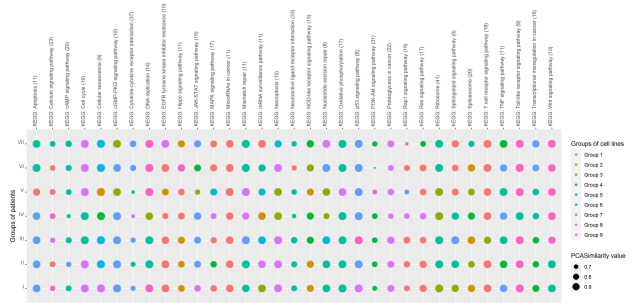
(a)



(b)



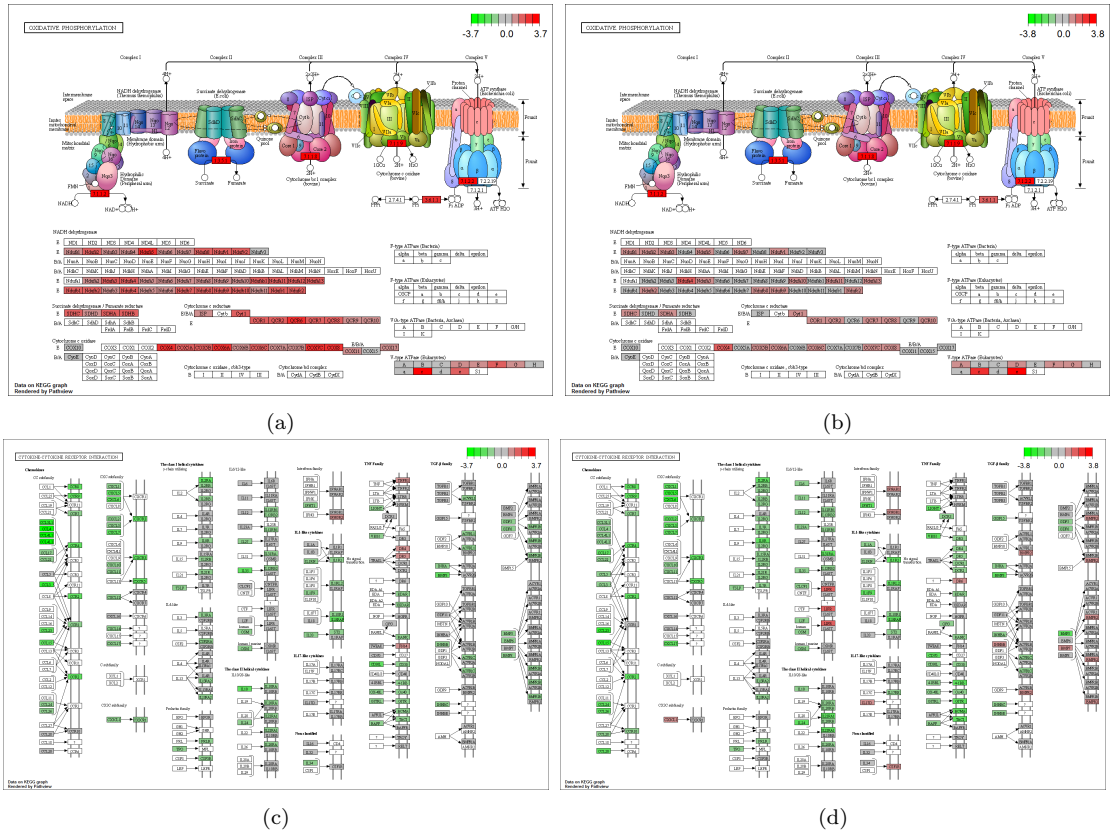
(c)



(d)

Figure 6: The similarity between groups of cell lines (1-9) and groups of tumors (I-VII). The size of the dot represents the degree of similarity between groups. The terms evaluated were (a) biological processes, (b) molecular function, (c) cellular component, and (d) metabolic and signaling pathways in KEGG.

F7 Example of similar/dissimilar expression profiles between groups of cell lines and groups of patients..



(a) (b)

(c) (d)

Figure 7: Example of similar and dissimilar metabolic pathways between group 5 cell lines and group VI patients. Gene expression in oxidative phosphorylation in (a) groups of cell lines and (b) groups of patients. Cytokine-cytokine receptor interaction gene expression in (c) groups of cell lines and (d) groups of patients.