ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

ԴԱՎԻԹԱՎՅԱՆ ՍՈՒՐԵՆ ՍԱՄՎԵԼԻ

ԿՐԾՔԱԳԵՂՁԻ ՔՂՑԿԵՂԻ ԵՎ ԳԼԻՈՄԱՅԻ ՄՈԼԵԿՈՒԼԱՅԻՆ ԲԱԶՄԱԶԱՆՈՒԹՅԱՆ ԲՆՈՒԹԱԳՐՈՒՄԸ ՏՐԱՆՍԿՐԻՊՏՈՄԱՅԻՆ, ԳԵՆՈՄԱՅԻՆ ԵՎ ԷՊԻԳԵՆԵՏԻԿԱԿԱՆ ՏՎՅԱԼՆԵՐԻ ՀԻՄԱՆ ՎՐԱ

Գ.00.02 - «Կենսաֆիզիկա, կենսաինֆորմատիկա» մասնագիտությամբ
կենսաբանական գիտությունների թեկնածուի գիտական աստիճանի
հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

ԵՐԵՎԱՆ – 2024

YEREVAN STATE UNIVERSITY

DAVITAVYAN SUREN SAMVEL

CHARACTERIZATION OF MOLECULAR DIVERSITY OF BREAST CANCER AND GLIOMA WITH TRANSCRIPTOMIC, GENOMIC, AND EPIGENETIC DATA

SYNOPSIS

of Dissertation Submitted for the Degree
of Candidate of Biological Sciences (Ph.D.) in the Field of
03.00.02. “Biophysics, Bioinformatics”

YEREVAN – 2024

Ատենախոսության թեման հաստատվել է Երևանի պետական համալսարանում:

Գիտական ղեկավար՝	կ.գ.դ. Արսեն Արտաշեսի Առաքելյան
Պաշտոնական ընդդիմախոսներ՝	ֆ.-մ.գ.դ. Արմեն Համլետի Պողոսյան կ.գ.թ. Արթուր Հրանտի Մուրադյան
Առաջատար կազմակերպություն՝	ՌԳԱ Սիբիրի բաժանմունքի Բջջաբանության և գենետիկայի ինստիտուտ

Ատենախոսության պաշտպանությունը տեղի կունենա 2024 թ. հունիսի 25-ին, ժամը 14:00-ին, Երևանի պետական համալսարանում գործող ՀՀ ԲԿԳԿ-ի Կենսաֆիզիկայի 051 մասնագիտական խորհրդի նիստում (0025, ք. Երևան, Ալեք Մանուկյան փ. 1, ԵՊՀ, կենսաբանության ֆակուլտետ):

Ատենախոսությանը կարելի է ծանոթանալ Երևանի պետական համալսարանի գրադարանում:

Ատենախոսության սեղմագիրն առաքվել է 2024 թ. մայիսի 22-ին:

051 մասնագիտական խորհրդի գիտական քարտուղար, կ.գ.դ., դոցենտ — Ս.Ա. Փարսադանյան

Dissertation topic approved at the Yerevan State University.

Scientific supervisor:	D.Sc. Arsen Artashes Arakelyan
Official opponents:	D.Sc. Armen Hamlet Poghosyan Ph.D. Arthur Hrant Muradyan
Leading organization:	Institute of Cytology and Genetics of the Siberian Branch of RAS

The defense of the dissertation will be held on 25th June 2024, at 14:00, at the session 051 Scientific Specialized Council on Biophysics of SCC of RA at Yerevan State University (0025, Yerevan, Alex Manoogian str. 1, YSU, Faculty of Biology).

The dissertation is available at the library of the Yerevan State University.

Synopsis has been sent on 22th May 2024.

Scientific secretary of 051 Specialized Council, D. Sc., Assoc. Prof. — M.A. Parsadanyan

INTRODUCTION

In recent years, numerous researches that examine the mechanisms of tumor development have led to the beginning of a new era of medicine (W. Liu et al., 2020; Martínez-Reyes & Chandel, 2021; Parker et al., 2020). Although these studies develop new treatment approaches (Chaturvedi et al., 2019; Cross & Burmester, 2006; Kamrani et al., 2023) and increase overall survival (DeSantis et al., 2011, 2014; Duffy, 2013) for many cancer types, there is a crucial need for comprehensive analysis to successfully predict, prevent, estimate survival rates, and treat patients. In parallel with the accumulation of new knowledge, there is growing recognition that cancer is difficult to study and treat due to its complexity and heterogeneity (Demicco et al., 2024).

The cancer research community predominantly identifies molecular diversity among various cancer types (inter-tumor heterogeneity) to enhance treatment accessibility to a broader population. Nevertheless, intra-tumor heterogeneity's significance is pivotal in personalized medicine and the scientific field as it contributes to examining novel characteristics in diverse subtypes.

Multiple alterations at transcriptomic, genetic, and epigenetic levels in many cancer types have been detected and described. Low-grade gliomas (Ozair et al., 2023) and breast cancers (Andrade De Oliveira et al., 2023; Sarhangi et al., 2022) are no exceptions as they showed differed expression, altered mutational burden, and methylation shifts in various processes and molecular functions. The studies emphasize the heterogeneity of low-grade gliomas (LGG) and breast cancers by examination of different subpopulations of cells within a tumor with differences in tumorigenicity, metastatic potentials, and therapy sensitivities (Haynes et al., 2017; Heppner & Miller, 1983; S. Wu et al., 2023). The high intra-tumor heterogeneity of these cancers suggests that potentially effective treatments could be missed if a specific molecular variation goes undetected. Moreover, the treatment strategy of patients correlates with molecular features that are affected at different levels. Therefore, there is a high necessity for the investigation of tools capable of considering peculiarities and conducting a comprehensive analysis of the molecular diversity of cancers. The application of such tools has already been widely practiced in scientific and clinical fields. While bioinformatic tools are mainly designed for observing peculiarities only at specific levels (Štancl & Karlić, 2023), machine learning-based approaches mostly solve the issues with the classification of samples or predict the type of cancer for patients with unknown diagnoses (Adams et al., 2023; Balkenende et al., 2022; Booth et al., 2020; Kadir & Gleeson, 2018; Kröner et al., 2021; Sultan et al., 2020).

Objectives and research tasks

The objective of this study is to develop and apply integrative multi-omic data analysis and knowledge transfer approaches to analyzing the molecular diversity of breast cancers and low-grade gliomas.

Research objectives are:

1. Develop a pipeline for integrative multi-omic analysis based on self-organizing maps approach;
2. Develop a transfer learning method for the projection and characterization of new samples on an existing self-organizing space;
3. Perform an integrated multi-omic characterization of breast cancer subtypes, evaluate the associations among functional, regulatory, and structural omic features, and examine their relationships with clinical indicators and prognosis.
4. Perform a multi-omic characterization to explore the molecular diversity of low-grade gliomas and evaluate the relationship between omic profiles and the disease's World Health Organization (WHO) genetic subtypes.

The scientific and practical significance of the results:

Breast cancer and low-grade glioma, examined in our thesis, demonstrate distinct molecular mechanisms that impact tumor prognosis and development. Identifying the underlying factors for the alteration in tumor samples serves as a basis for targeted and personalized treatments customized for particular subtypes and individual samples. Although an array of studies and models have been developed and validated for analyzing various tumors, they are often limited to general examination because of the heterogeneity and complexity of cancer. As mentioned above, a multi-omics-based machine learning approach in our analysis allows for addressing these problems. Namely, this method is designed to reveal the molecular diversity of cancers connecting the altered factors seen in samples and the pathological traits observed in clinical analysis. The integrative pipeline developed in the thesis facilitates the identification of disturbed gene modules across diverse omic landscapes. The model demonstrated significant results for both types of cancer, which provides evidence for the validity of the proposed method.

Our results showed alterations on the transcriptome layer for processes associated with proliferation, epithelial-mesenchymal transition (EMT), immune response, DNA repair, and stromal/stem cell signature for breast cancer PAM50 subtypes. The luminal A subtype showed significant differences in comparison with other subtypes, which can be related to the more aggressive nature and worse prognosis of luminal A cancers. The results also highlight subtype-specific associations of the same transcriptomic alterations and methylation or CNVs or SNVs. For the basal subtype, we observed the highest values for the expression of EMT genes and hypomethylation of the mentioned signature. Additionally, EMT genes were overexpressed in luminal A and luminal B cancers and showed a positive correlation with CNV counts. For low-grade gliomas (LGG), we demonstrated that the classification of LGG samples according to the expression and methylation values is more informative from the prognostic point compared to the genetic subtypes.

Finally, our results emphasize the complex subtype-characteristic associations between gene expression and epigenetic/genomic factors and their implications for survival and clinical outcomes.

Approbation. Proceedings of the thesis have been presented at "Lomonosov 2021" May 2021, Moscow; "Lomonosov 2022", April 2021, Moscow; "16th RAU annual conference" April 2021, Yerevan; "17th RAU annual conference" March 2022, Yerevan; "International Congress on Informatics: Information Systems and Technologies" November 2022, Belarus; "Society of Immunotherapy of Cancer 38th Annual Conference", March 2023, San Diego; "Society of Immunotherapy of Cancer 39th Annual Conference", April 2024; Houston.

Publications. The main results of the dissertation are published in 3 papers and 2 presentation abstracts at international scientific conferences.

Structure. This dissertation comprises 100 pages of computer-formatted English text, including 2 tables and 29 figures, consisting of the following sections: Introduction, Literature Review, Materials and Methods, Results and Discussion, Conclusion, Inferences, References (including 266 sources), and Appendix (pp. 101-147).

LITERATURE REVIEW

The literature review examines the variety of molecular mechanisms of breast cancer and gliomas and emphasizes the heterogeneity of diseases. The availability of sources of multi-omic data is discussed, various machine learning algorithms are mentioned, their advantages and disadvantages are emphasized, as well as their applicability in research of multi-omic biological data.

MATERIALS AND METHODS

Breast cancer data. In this study, we used available omic datasets of the TCGA-BRCA (Breast Invasive Carcinoma) project (The Cancer Genome Atlas Research Network et al., 2013). RNA-seq counts, microarray promoter and gene body methylation, CNV, and SNV were obtained for 996 samples. The patient's clinical data was retrieved from the GDC database and contains variables such as tumor pathologic stage, information about treatment, survival, etc. We used gdc-client (Grossman et al., 2016) to download the required files from the portal. PAM50 molecular classification data for these samples were obtained from the publication by Chia et al. (Chia et al., 2012).

STAR-aligned and TPM-transformed expression values for each sample were obtained. Expression values were then subjected to library size normalization and converted to log2(counts+1) using a variance stabilizing transformation.

The GDC-supplied methylation data contains betta values (Methylation intensity/(Methylation intensity+Unmethylation intensity)) measured with Illumina Infinium DNA Human Methylation 27 and HumanMethylation 450 microarrays. DNA methylation data was aggregated by merging the data generated with two microarrays. To obtain per gene promoter level methylation data, the betta values of CpG islands that overlap with the corresponding promoter were averaged. Next average promoter beta value for each gene was converted to methylation M values as follows: M = log2(Beta/1-Beta) (Du et al., 2010).

CNV available for the project was generated with the Affymetrix SNP 6.0 genotyping platform. We used ASCAT (Van Loo et al., 2010) to produce gene-level copy numbers. These values were further log2 transformed. Prior to normalization, we added small random numbers (mean=0, SD=0.001) to the CNV data to row-wise constant values.

SNV data was obtained from the GDC portal in the form of Mutation Annotation Format (MAF) files. MAF contains aggregated mutation information from VCF files and is produced from Somatic Aggregation Workflow. In MAF files multiple SNVs for the same gene, therefore, we summarized SNV counts per gene. Lastly, and similarly to CNV data, we added small random numbers row-wise SNV counts before normalization to avoid constant values in the final data.

Low-Grade Glioma data. The glioma dataset consisted of 122 World Health Organization (WHO) grade II and III adult-type glioma (low-grade glioma, LGG) samples collected in the framework of the German Glioma Network (GGN) (Weller et al., 2009). In addition, we collected IDH-A gliomas with single Chr19q deletions without Chr1p co-deletions into a separate class IDH-A (astrocytoma).

The gene expression of GGN glioma cohort samples was measured using Affymetrix Human Genome U133 Plus 2.0 microarrays. Methylation was measured using Illumina HumanMethylation450 BeadChip arrays and presented as M values. CNV levels were obtained using array-CGH microarrays. Gene expression and methylation data are available in the Gene Expression Omnibus (GEO) database under accession numbers GSE61374 (LGG expression (Weller et al., 2015)) and GSE129477 (LGG methylation (Binder et al., 2019)).

Integrated analysis of cancer molecular features with multi-layer SOM. To conduct an integrative analysis of omic datasets of breast cancer, we developed a multilayer self-organizing maps (ml-SOM) approach as an extension of the oposSOM package described in detail in the previous section and elsewhere (Binder et al., 2022; Loeffler-Wirth et al., 2022). For ml-SOM, we organized all omic datasets into distinct layers and trained them collectively on a single SOM grid, similar to a classical single-layer SOM (sl-SOM) of the oposSOM package (Löffler-Wirth et al., 2015). The key distinction between the training of sl-SOM and ml-SOM lies in how the best matching unit (BMU) is selected within the SOM grid.

In sl-SOM, the BMU is chosen based on the distance between the input vectors and the weight vectors of SOM nodes (Wirth et al., 2012). However, in the case of ml-SOM, these distances are calculated separately for each layer and then combined into a single value, taking into account the respective layer weights as follows:

$$D = \sum_{i}^{n} \omega_i * d_i$$

where n - number of layers, i - weight of ithlayer, di - distance to the SOM node on ith layer.

The weight factor ω scales the effect of each of the layers on the topology of the ml-SOM. It takes values from 0 to 1 and ensures that $\sum_{i=1}^{n} \omega_i = 1$. Because SOM training applies to the combined multimodal vectors the topology of the resulting map is governed by the weighting factors, which, in turn, define the degree of couplings between the different omics layers. For the breast cancer dataset, we applied ωGex = 1, ωGmx = 0, ωCNV=0, and ωSNV=0 weights to force arrangements of genes on the SOM grid by expression data. For the glioma dataset, we used equal weights ωGex = ωGmx = ωCNV=1/3 to ensure the balanced coupling between expression, methylation, and CNV layers.

The downstream analysis of ml-SOM is similar to the oposSOM pipeline (Löffler-Wirth et al., 2015). Due to the self-organizing properties of the SOM, neighboring nodes tend to have similar weight vector profiles, which can be visualized as a SOM portrait by applying a color gradient (for example, from blue to red). As a consequence, the obtained mosaic images show a smooth texture with red and blue spot-like regions referring to clusters of increased or decreased omics scores in the respective sample or sample group. Following the Self-Organizing Map (SOM) training, we partitioned the resulting metagene map into discrete regions referred to as "spots." These spots represent clusters of genes that exhibit similar co-(de)regulation patterns, particularly genes that are perturbed in at least one of the studied omic layers. Spot dissection (selection of genes) within SOM portraits can be performed using various criteria, such as overexpression (for glioma), 'k-means,' and 'variance' (breast cancer) (Wirth et al., 2012).

For biological function mining in these spots, we performed functional annotation using overrepresentation and Gene Set Z-score analyses. Additionally, we also employed the Enrichr resource (Kuleshov et al., 2016) for over-representation analysis using additional gene sets.

Tumor Similarity Analysis, Supporting and Prognostic Maps. Similarity analysis compares the SOM portraits of the tumor samples by means of Pearson's correlation coefficient of their omic layer values using meta-genes instead of single genes, which improves the representativeness and resolution of the results (Hopp et al., 2013). The correlation matrix was visualized using pairwise correlation maps and correlation net representations. The correlation net constructs an unweighted graph by connecting the nodes (samples) whose pairwise correlation coefficient exceeds a given threshold ($r > 0.5$).

Association analysis of the association between molecular features in tumors. To assess the relationships between gene expression, methylation, CNV, and SNV across various spots (gene modules), we utilized linear regression with gene expression as the outcome and the other genomic markers as explanatory variables. The model was further enhanced by incorporating cancer subtypes as an interaction term, allowing us to examine the variability in these relationships across different disease subtypes and omic layers.

We used the 'emmeans' (Searle et al., 1980) and 'interactions' (Bauer & Curran, 2005) packages in R were employed for statistical analysis and visualization of these interactions. Additionally, we applied the Dunnett's Test to evaluate pairwise differences in mean levels of expression and methylation, CNV, and SNV gene modules in subgroups compared to true normal tissue. Another approach for analyzing omic layer associations was to calculate the signed square root covariance (ScoV) between omic metagene profiles in a pairwise fashion (Binder et al., 2022). For the sample groups, we calculated mean portraits by averaging the respective metagene values for the overall individual sample portraits of the respective group.

Creating "phenotype" maps of association with clinical characteristics and survival. Phenotype data accompanying the breast cancer dataset, such as medication, disease stage, and grade for TCGA-BRCA samples was obtained from the GDC portal. Phenotype data of the glioma dataset was obtained from the German Glioma Network.

To create a phenotype map, we constructed a linear regression model of metagene value as a dependent variable and clinical parameters as an ordinal independent variable. Then, we mapped the corresponding regression coefficient for the predictor variable to a corresponding position of a metagene on the SOM grid. The visualization of weight coefficients allows evaluation of the association of corresponding clinical characteristics and the levels of functional gene modules on different omic layers.

Survival analysis was performed using the Cox proportional hazards regression using 'contsurvplot', 'survival,' and 'survminer' R packages. The model includes survival as a dependent variable and spot omic profiles and group information as predictors.

Projection of New Samples Into Existing SOM space. Supervised SOM (supSOM) portrayal is based on support vector machine regression (SVMR) and provides an alternative approach for extending an existing SOM space. In supSOM, one SVMR model is trained for each meta-gene individually, using the genes' expression profiles of the primary data as an independent variable and the corresponding meta-gene profile obtained from the initial SOM training as the dependent variable. Thereby, only genes associated with the particular meta-gene or one of the adjacent meta-genes are considered predictors. Once a model is trained, gene profiles in new samples can be used to predict the corresponding meta-genes. We applied the SVM regression model with the Gaussian kernel and evaluated supSOM performance for varying neighborhood radii.

Performance and accuracy for supSOM were assessed based on the evaluation of correlation and root-mean-square deviation (RMSD) between metadata of the extension samples (i.e., the portraits) generated by SOM as reference vs. supSOM.

For benchmarking runtime of the SOM initialization and training phases, we generated artificial expression matrices for the primary and secondary (extension) data (m1 = m2 = 50, 100, 200, 500, and 1000 arrays per class) using the 'madsim' R package (Dembélé, 2013). As a "real life" data use case, we studied disease grade-associated transcriptome changes in breast cancer datasets (GEO accessions: GSE42568, GSE10810, and GSE29431).

RESULTS AND DISCUSSION

Extending the functionality of the SOM pipeline. We developed a new SOM cartography method to perform integrative analysis and visualization of molecular landscapes. Our method dissects the different omics landscapes into modules of co-methylated, co-expressed, and co-aberrant genes focused around a particular biological process or function. They reflect the underlying network of regulatory modes of cell activity within each of the omics layers and between them. With our ml-SOM pipeline, it is possible to combine as much omic data as available as long as they can be represented in a gene-centric way. In the breast cancer dataset, we were interested in expression-driven clustering of gene modules in PAM50 subtypes, so we used heavily expression-imbalanced weights. Contrarily, equal weights for expression, methylation, and CNV layers were applied in the analysis of the glioma dataset to identify molecular subtypes and associate them with known genetic classes (IDH-mut, IDH-wt).

The ml-SOM approach offers another important addon to the integrative analysis of multi-modal omic data. It is known that the transcriptome and proteome are dynamic and reflect the functional state of a cell (Unwin & Whetton, 2006). They can provide significant insights into the molecular mechanisms of cancer development and progression. On the other side, epigenetics and genomic aberrations significantly impact the dynamic state of the cell (Ducasse & Brown, 2006; Sadikovic et al., 2008). The ml-SOM also opens an opportunity to evaluate the association between dynamic (transcriptome or proteome) and regulatory (methylation, CNVs, and SNVs) omic layers using regression of covariance.

Finally, multi-omics cartography in terms of phenotype maps provides a tool to extract gene signatures associated with clinical indicators and survival across different omics layers.

In conclusion, ml-SOM cartography allows for disentangling the diversity of regulatory modes of cell functions in terms of easy-to-interpret gene-centric data landscapes. Due to the growing use of multi-omics data, we expect these options will become important for future progress in cancer bioinformatics.

Integration of new samples into SOM space. The major disadvantage of the SOM method is its inability to integrate new samples into an existing SOM space without retraining the entire model. This can be time- and computing resource-consuming; moreover, due to the initialization phase of the SOM, the arrangements of the genes on a grid can be different and incomparable. The same applies to the ml-SOM pipeline. To overcome these drawbacks, we developed a supervised approach (supSOM) that adds a support vector machine regression model on top of the original SOM algorithm and "predicts" the SOM portrait of a new sample. The general workflow of supSOM is presented in Figure 1.

The "primary" dataset is trained with self-organizing maps (SOM), followed by clustering and downstream analysis. Then, the support vector machine regression model (SVMR) is trained to map the input dataset to SOM "portraits" generated from "primary" data. Finally, a "secondary" dataset is supplied to the model for projection into the SOM space. In supSOM, one SVMR model is trained for each meta-gene individually, using the genes' expression profiles of the primary data as an independent variable, and the corresponding meta-gene profile obtained from the initial SOM training as a dependent variable. Only genes associated with the particular meta-gene or one of the adjacent meta-genes of defined radius are considered predictors.

Once a model is trained, gene profiles in new samples can be used to predict the corresponding meta-genes. The accuracy of prediction based on the simulated dataset was 0.9-0.99 depending on the radius. We successfully tested supSOM for analysis and prediction of breast cancer histologic grades. The SOM was trained with a "primary dataset" (GSE42568) that contains gene expression profiles measured in 121 samples stratified by breast cancer histologic grade (17 normal, 11 Grade

I, 40 Grade II, 53 Grade III). The supSOM analysis of "secondary" datasets GSE29431 and GSE10810 showed a good correlation with primary SOM counterparts (Figures 2A and 2B).

In conclusion, the supSOM is a transfer learning approach that projects novel data into a multidimensional space obtained from previously collected data. It considerably widens the application range of SOM portrayal by reducing model limitations and computation demand and usage of previously generated knowledge for the characterization of new samples.

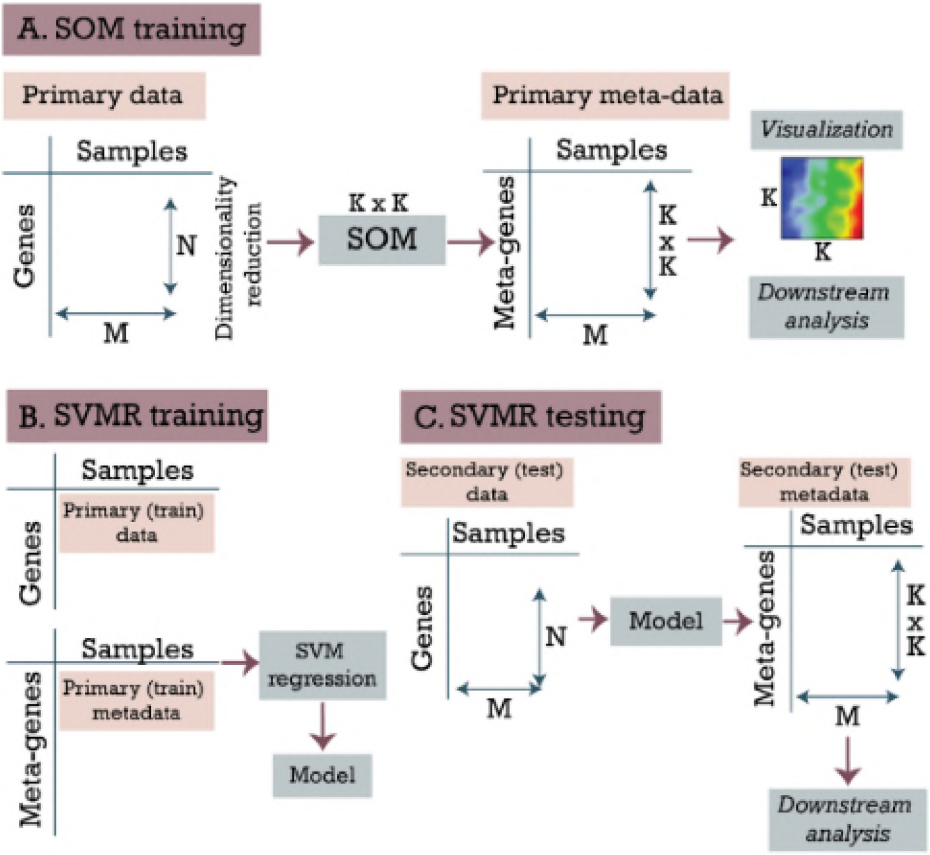

Figure 1. General workflow of the supSOM algorithm. In supSOM, the SVMR model is trained to map the primary dataset to its SOM "portraits." During supSOM testing, the "secondary" dataset is supplied to the model for projection into the SOM space. Single arrows indicate the order in the pipeline, while double arrows the dimensions of samples/features in the matrix.

Integrated analysis of omic landscapes in breast cancer subtypes. The TCGA-BRCA multi-omics data that contains information about 996 samples were classified by PAM50 molecular classification and were analyzed by multi-SOM algorithms (Figure 3). The expression, promoter methylation, CNV (copy number variations), and SNV (single nucleotide variants) data were analyzed with the ml-SOM algorithm, which was trained on a 40x40 node grid for dimension reduction.

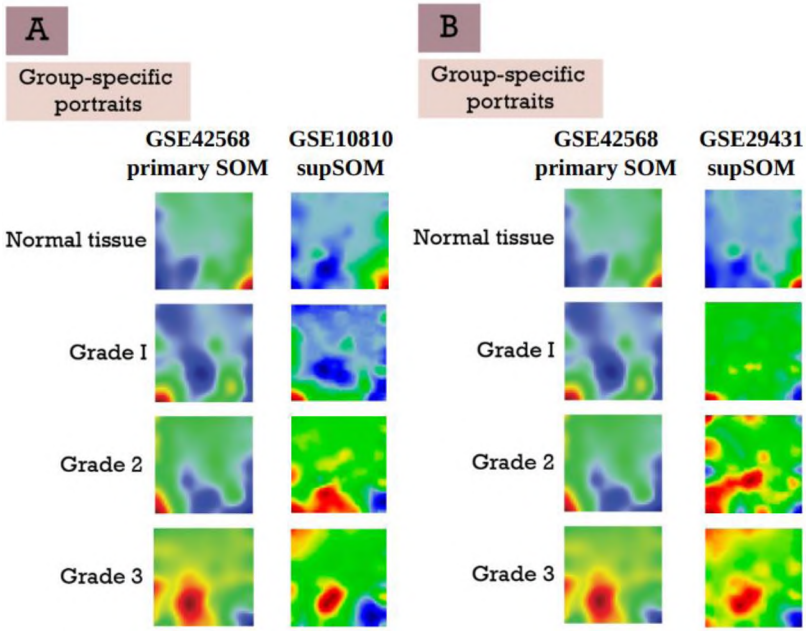

Figure 2. Comparison of supSOM portrayal of breast cancer transcriptome landscapes. (A) Portrayal of GSE42568 (primary) and GSE10810 secondary/extension). (B) Portrayal of GSE42568 (primary) and GSE29431 (extension).

During the training phase, ml-SOM combines the genes having similar profiles of expression, methylation, CNV, and SNV across samples into adjacent nodes according to the weight factors on the SOM grid thus forming gene clusters (also referred to as spots or gene modules). We combined samples in groups for downstream analysis according to the PAM50 classification to assign cancer samples to molecular subtypes. The average multi-omic SOM portraits showed considerable variations in expression (Gex), methylation (Gmx), CNV, and SNV both across PAM50 subgroups as well as compared to true normal tissue. The ml-SOM generated variance landscapes were further segmented into gene modules (or spots) with 'k-means' criteria and highly variant spots (spots A, C, E, F, L, R, Q, S) were selected for downstream analyses. The average number of genes per spot was 155±60 (M±SD). Spot F had the lowest gene counts (9 genes), while spot R had the highest (207 genes). Next, we performed a downstream functional analysis with Gene Set Enrichment Analysis (Löffler-Wirth et al., 2015) and an over-representation analysis with genesets covering multiple domains (Kuleshov et al., 2016).

Spot A contained 114 genes, associated primarily with DNA replication (padj = 3.3e-05), E2F targets (padj = 5.3e-09), retinoblastoma pathway (padj = 3.23e-05), and cell cycle activity (padj = 5.4-e04). Notably, spot genes were also associated with EMT markers taken from Sarrió et al (Sarrió et al., 2008), as well as markers for the basal BC subtype taken from Smid et al (Smid et al., 2008).

Spot C contained 165 genes mostly involved in protein transport (padj = 2.0e-02), SMARCA2 antiproliferative targets (padj = 2.2e-06) (Shen et al., 2008), and DNA repair. (padj = 4.34E-03).

Spot E contained 118 genes enriched with luminal cancer gene signatures (padj = 0.005) (Charafe-Jauffret et al., 2006) and genes associated with the amplification of chromosome 16p13 (padj = 0.02) (Nikolsky et al., 2008).

Spot F contained only nine genes; however, they were implicated in vitamin D signaling (padj = 0.038), palmitoyl-CoA Hydrolase Activity (padj = 0.025), Androgen Receptor/NKX3-1 Signaling (padj = 0.01) and ICGC transcription factor target genes (padj < 0.01).

Spot L contained 67 genes strongly associated with the immune system process (padj = 1.2e-12).

Spot Q contained 141 genes enriched with stromal (padj = 7.84E-03)(Finak et al., 2008) and stem cell gene signatures (padj = 8.81E-14) (Lim et al., 2010), genes involved in accelerated proliferation (padj = 6.1e-05), inflammation (padj = 0.03), RAS signaling (padj = 1.1e-05), hypermethylation of tumor suppressor genes (padj = 3.2e-04).

Spot R contained 207 genes associated with RNA splicing (padj = 0.005) and mitochondrial gene signatures (padj = 0.004).

Spot S contained 106 genes enriched with luminal cancer signatures (padj = 4.3e-06), ESR1 signatures (padj = 3.5e-15), and metastasis-suppressing signatures (padj = 2.1e-03). We labeled the selected spots based on the functional annotations that best describe the genes associated with each spot (Table 1).

Table 1. Spot function assignment along with the top 3 correlated genes.

Spot	**Top correlated genes on transcriptome SOM layer (Pearson correlation coefficient, r)**	**Spot Assignment**
A	*RAD51AP1* (r = 0.81), *KIF2C* (r = 0.76), *DSCC1* (r = 0.76)	cell cycle, metastasis, EMT
C	*HLTF* (r = 0.83), *GIT2* (r = 0.80), *ACAP2* (r = 0.80)	miRNA targets/DNA repair
E	*ROGDI* (r = 0.77), *RAB26* (r = 0.75), *HAGH* (r = 0.75)	luminal cancer
F	*TATDN3* (r = 0.64), *THEM4* (r = 0.62), *DCAF8* (r = 0.58)	VDR signaling
L	*FERMT3* (r = 0.93), *PARVG* (r = 0.86), *FMNL1* (r = 0.88)	immune response
Q	*CAV1* (r = 0.81), *TGFBR2* (r = 0.77), *RBMS1* (r = 0.75)	stroma/stem cells
R	*SSNA1* (r = 0.85), *DRAP1* (r = 0.82), *SURF2* (r = 0.82)	RNA splicing
S	*SCUBE2* (r = 0.82), *ESR1* (r = 0.82), *ABCC8* (r = 0.76)	ESR1 signaling

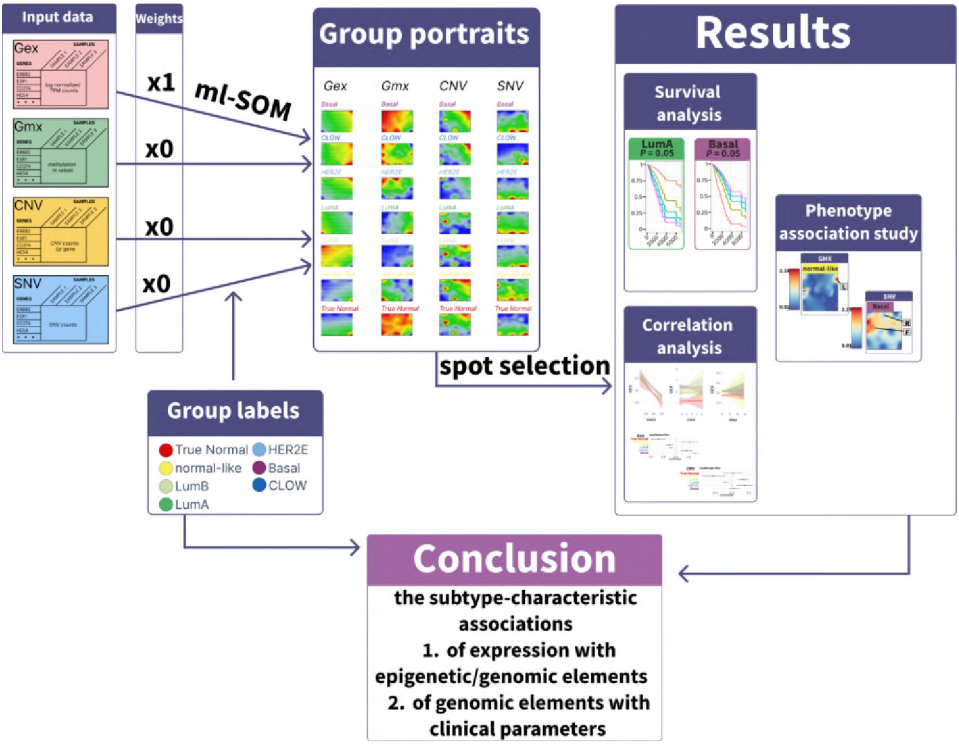

Figure 3. Schematic summary of the multi-omic analysis of breast cancer PAM50 subtypes.

Multi-omic summary of deregulated modules in breast cancer subtypes, survival, and clinical phenotypes. We aimed to summarize findings from multi-omic analyses based on breast cancer subtypes, focusing on gene modules, survival, and phenotypic characteristics. For this purpose, we constructed Cox regression models for the interaction of continuous expression, methylation, CNV, and SNV levels for each gene module and each cancer subgroup. We also generated phenotype maps visualizing the association between clinical phenotype parameters with different omic layers as described in our previous publication (Arakelyan et al., 2021).

We found that gene signatures associated with EMT/cell cycle, luminal, immune system, and RNA splicing were upregulated across all cancer subtypes compared to normal tissue. Conversely, stromal/stem cell signatures were downregulated across all cancer subtypes. The expression levels of RNA splicing genes remained consistent across all cancer subtypes. Immune signature genes were notably higher in HER2E, basal, and normal-like cancers than in luminal A and B subtypes. For other gene modules, the extent of expression was varied along with cancer subtypes. Specifically, the expression of the EMT/cell cycle module progressively increased from luminal A through normal-like, luminal B, HER2E, to basal cancers, with the highest expression noted in basal cancers. Similarly, for luminal gene signature, the expression gradually increased from basal through normal-like, HER2E, luminal A to luminal B cancers. Finally, gradual downregulation of stromal/stem cell signature was observed from normal-like through basal, luminal A, HER2E to luminal B subtypes. Interestingly, these changes were paralleled with the increase in methylation

levels. In addition, we observed consistently increased methylation levels of VDR genes across all cancer types, except the normal-like category. However, this was not associated with any noticeable changes in their expression levels. In addition to these changes shared by all cancer subtypes, there were more subtype-characteristic perturbations.

Thus, luminal A cancers were additionally characterized by downregulation of expression and methylation of DNA repair genes, and overexpression of ESR1 signature genes as the most dominant feature for this cancer subtype. Finally, this subtype showed decreased counts of SNVs in the immune system, stromal/stem cells, and RNA splicing genes. Interestingly, EMT/cell cycle gene expression in this subgroup was upregulated despite increased methylation levels; however, their expression levels were positively correlated with CNVs. The survival in luminal A cancers was associated with several gene modules on different omic layers, with the highest impact of the low expression levels of EMT and DNA repair genes on favorable survival prognosis. The luminal A cancers showed multiple significant associations with clinical phenotypes (Figure 14). Particularly, the overexpression of DNA repair genes was associated with poor prognosis. Moreover, increased SNV profiles and decreased methylation of these genes in this cancer type were associated with advanced stages of American Joint Committee on Cancer's (AJCC) tumor pathologic assessment; particularly with pathologic M (metastasis) and pathologic N (lymph nodes). Furthermore, the increased expression of luminal cancer gene signature was associated with the presence of prior malignancies.

The luminal B subtype exhibited expression changes similar to luminal A cancers, except for unchanged expression DNA repair genes and a downregulated ESR1 signaling gene signature (spot S). This specific downregulation in luminal B was not linked to significant changes in methylation or CNV when compared to luminal A cancers. However, it showed a negative correlation with the SNV profile, not observed in luminal A cancers. Moreover, the methylation profiles in these two luminal subtypes closely resemble, except for increased methylation of immune system genes in the luminal B cancers. We did not observe any significant association with survival in this cancer subtype, except for methylation of immune system genes with borderline significance ($p=0.0678$) (Figure 23). Phenotype portraits for this cancer subtype showed a positive association of advanced AJCC pathologic staging, and, in particular AJCC pathologic M with expression and methylation of RNA splicing genes as well as decreased expression and methylation of DNA repair genes. AJCC pathologic T (primary tumor) was positively associated with the increased CNV profiles of the EMT/cell cycle and immune response genes.

The transcriptome profiles for the HER2E subtype were closely aligned with those of luminal B, differing only in the magnitude of changes. Unlike in the luminal B subtype, methylation levels of EMT/cell cycle genes in the HER2E subtype remained unchanged compared to the true normal samples. Furthermore, this subtype exhibited increased CNV profiles for luminal and stromal/stem cell gene signatures. Notably, the survival impact for this cancer subtype was most influenced by the underexpression of the VDR gene signature and the overexpression of immune system genes (Figure 4).

In agreement with the survival data, the negative association of vital status with the overexpression of immune system genes was observed. Moreover, the overexpression of the luminal cancer signature was positively correlated with advanced tumor staging (Figure 5).

Unlike Lum A, Lum B, and HER2E cancers, in basal cancers, a strong negative correlation existed between the overexpression of EMT/cell cycle gene signatures and decreased methylation levels. A similar relationship between underexpression and methylation levels was noted for DNA repair genes. This cancer subtype's survival was linked to various gene modules across omic layers, with hypomethylation of the ESR1 gene signature having the most significant positive impact on prognosis (Figure 4).

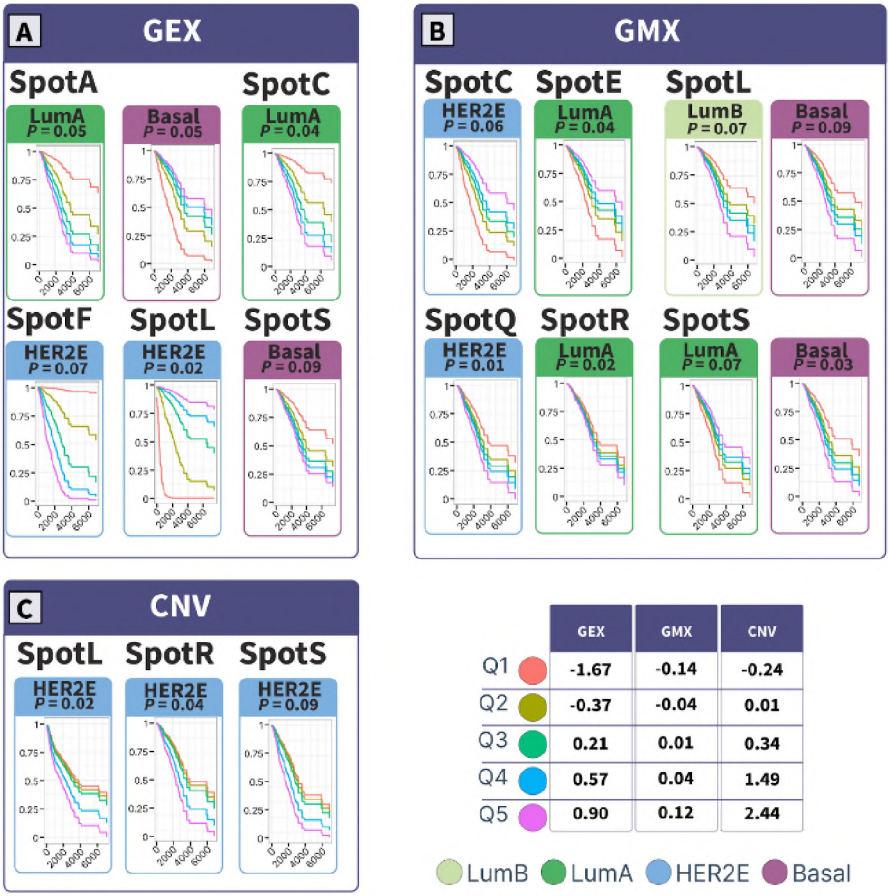

Figure 4. Association of omic features of SOM gene modules with survival in PAM50 subtypes. Survival analysis was performed using the Cox proportional hazards regression with the inclusion of spot levels as continuous variables using "contsurvplot," "survival," and "survminer" packages. Survival curves were visualized with range values with 5 intervals (Q1: minimum, Q2: 25th percentile, Q3: 50th percentile, Q4: 75th percentile, Q5: maximum). Only plots with survival-spot association with p-value ≤ 0.1 are displayed.

Furthermore, the prior malignancy was positively associated with increased methylation of DNA repair genes and decreased methylation of RNA splicing genes (Figure 5). Finally, the normal-like cancers were characterized with additional underexpression of VDR signaling signature and a decrease of CNV profiles in almost all gene modules (DNA repair, luminal cancer, VDR, immune response, stromal/stem cells, RNA splicing) compared to other cancer subtypes. No specific gene modules were significantly associated with survival for this cancer subtype, however, we observed a significant association of CVN increase in VDR signaling, EMT/cell cycle, and immune system genes with prior malignancy in this cancer subtype (Figure 5).

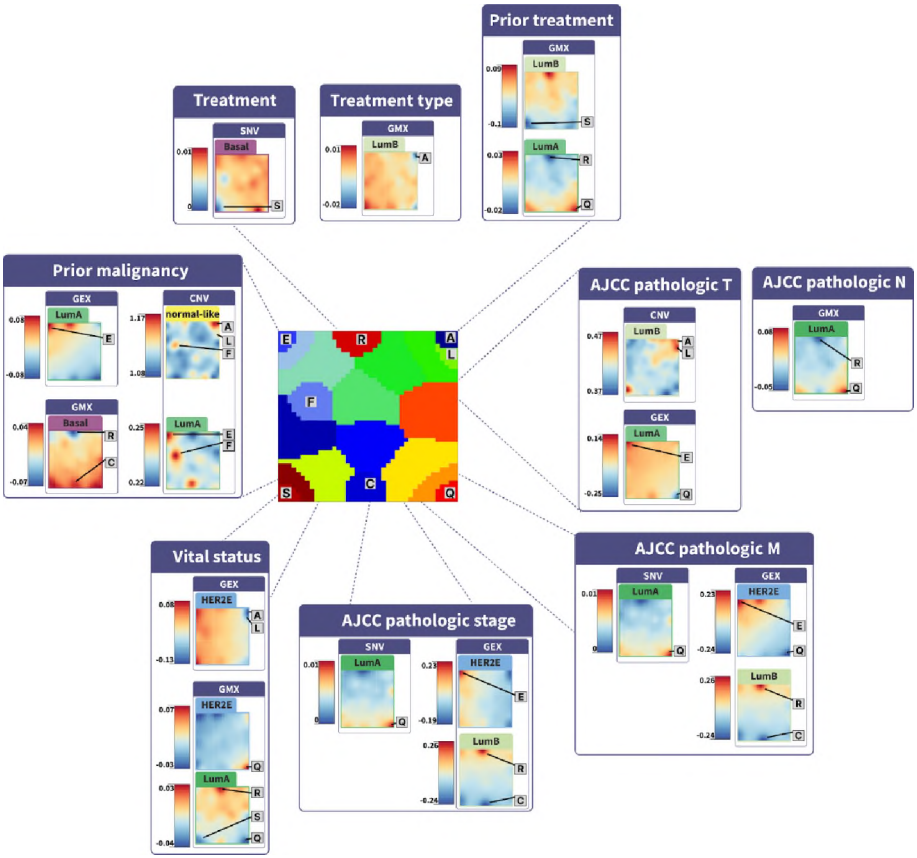

Figure 5. The phenotype portraits of the association of omic features of SOM gene modules (spots) with clinical parameters in PAM50 subtypes. Phenotype portraits show the -log10(p) regression model of SOM metagene levels and clinicopathological stages, vital status, and treatment variables. Deregulated gene modules (spots) are indicated on the maps, and coloring shows the significance of their association with evaluated parameters.

Integrated Multi-Omics Cartography of Lower-Grade Gliomas. The Low-Grade Glioma cases were classified into four subtypes following WHO 2021 records(Louis et al., 2021). The unmutated (wild type) IDH1 and/or IDH2 (IDH) gliomas were classified as IDH-wt, meanwhile, LGG with IDH gene mutations were divided into the IDH-O (Chr1/19codel), IDH-A (astrocytomas without Chr1/19codel) and IDH-A' (IDH-mut astrocytomas with Chr1/19codel). The same LGG samples were previously classified into 8 transcriptomic E1-E8 and 6 methylation M1-M6 groups based on their -omics characteristics (Binder et al., 2019; Willscher et al., 2021). The E1 and M1 groups belong to IDH-wt, E6, and M5 in IDH-O, and E2–E4 and M2–M4 in IDH-A/A'. The E7, E8, and M6 subtypes belong to the neuronal (NL) tumors and have decreased cancer cell content. The ml-SOM algorithms were used in three-omics data that contains transcriptomic

(Gex), methylation (Gmx), and CNV layers. Consequently, group portraits are portrayed for each of the layers. The associations between layers and subtypes were observed by creating pairwise correlation heatmaps. E7, E8, and M2 were related to NL tumors, subtypes M2 and E3 have patterns in the IDH-A group due to their decreased methylation level forming a separate methylator type (Figure 6).

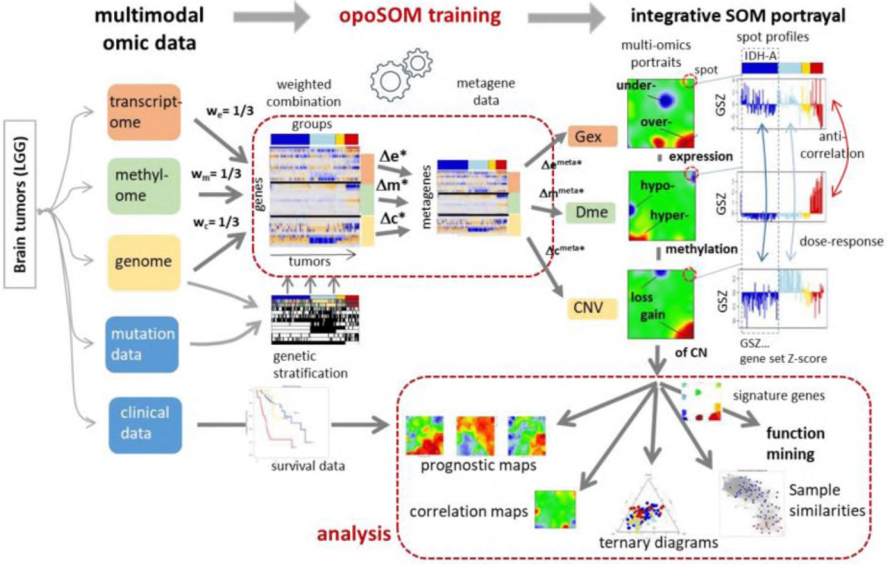

Figure 6. Schematic summary of the multi-omic analysis of low-grade gliomas.

For the subsequent stages of our research, we have generated group-averaged Gex, Gmx, and CNV portraits. The ml-SOM algorithm specified 11 spots that were labeled with uppercase letters (A-K). The spots A, B, C, and G were altered in Gex and Gmx (green), D, E, F, H, I, and J for CNV and spot K only in the Gex layer. Between Gex and Gmx selected spots (spots B and G), we have observed a negative correlation, in contrast to Gex and CNV (spots E and J) which have shown positive correlations. Gene set analysis showed that the functional context of these spots is relegated to healthy brain, proneural, inflammation, and EMT (epithelial-mesenchymal transition) transcriptional signatures in the Gex map, to targets of the polycomb repressive complex 2 (PRC2) related to neural development and to GCIMP, GCIMP-O, GPCR, and RTKII (receptor tyrosine kinase type II methylation subtype) methylation modes in the Gmx map and to individual chromosomes (mostly chromosomes 1, 7, 10, and 19) in the CNV. Notably, Gmx and CNV changes were not correlated, but both layers were correlated to the expression layer.

We further decided to compare the genetic subtypes with the E- and M- groups across the three -omics layers to identify factors that are associated with changes in layers of multi-omics data. The transcriptomic, methylation, and copy number portraits of the IDH-wildtype subtype were similar to the E1 and M1 groups, while the E6 and M5 subtypes were closer to the IDH-O subtype. The IDH-A group was highly similar to E4 and M4 groups, and the NL-like subtype was distributed across all genetic groups because of their low cancer cell content. These results highlight the fact that the LGG heterogeneity can be better described in terms of transcriptomic- and epigenetic-based subtyping rather than by genetic groups.

CONCLUSIONS AND INFERENCES

The explosion of big biological data has opened unprecedented opportunities for the application of machine learning methods for understanding disease biology, patients stratification and developing efficient targeted/personalized treatments. From many different ML pipelines applied to the multi-dimensional omic data, the self-organized maps stand out for their unprecedented capabilities of dimension reduction with minimal information loss, feature extraction, and visualization. The "SOM portrayal" method has been applied for solving a wide spectrum of biological questions, such as understanding of molecular basis of cancers, neurodegeneration, and aging, identification of disease molecular subtypes, performing patients stratification and linking it to disease prognosis, etc. Meanwhile, SOM method usage pointed to its major drawbacks: inability to integrate diverse omic data and issues with extending with new samples without retraining. In response to these issues in this thesis, we developed two approaches each handling the issue described above. Out multi-layer SOM pipeline can perform integrative analysis of gene-centered multi-omic data (expression, methylation, CNV, and SNV), and extract perturbed gene modules across different omic domains. Moreover, ml-SOM offers a new layer of analyses aimed at understanding the mutual relations between functional (gene expression), regulatory (methylation), and structural (CNV and SNV) data, to understand the contribution of each domain on the functional state of a cell. Finally, the capability to relate the changes on the molecular levels with clinical indicators and survival may lead to more personalized and effective treatment strategies.

Another significant improvement over existing SOM portrayal is the development of transfer learning approach (supSOM) to project new samples to available SOM space. This is particularly important since it preserves the previous knowledge making comparisons and interpretations "context"-based.

Cancers have greatly benefited from the advance of ML in the analysis of biological data. The complexity of cancers and the diversity of perturbations on different omic layers call for the development of integrated data analysis approaches. Here we demonstrated that our ml-SOM approach can handle this and provide valuable insights into disease molecular diversity, mechanisms of development and progression, and extract important prognostic indicators.

In breast cancer we comprehensively characterized the perturbed gene modules and biological processes in PAM50 subtypes. Our results showed mostly qualitative changes on the transcriptome layer for processes associated with proliferation, epithelial-mesenchymal transition (EMT), immune response, DNA repair, and stromal/stem cell signature. We also observed a principal difference in the expression of estrogen receptor signaling genes between luminal A and other cancer subtypes, which can explain the more aggressive nature and worse prognosis in the latter. Moreover, we demonstrated that the same expression perturbations may be associated with methylation or CNVs or even SNV in subtype characteristic manner. As an example, the expression of EMT genes were highest in basal cancer and was associated with hypomethylation, while the modest overexpression of the same genes in luminal A and luminal B cancers were positively associated with CNV counts. Finally, pur results highlight the complex subtype-characteristic associations between gene expression and epigenetic/genomic factors and their implications for survival and clinical outcomes.

The low-grade glioma is another cancer type characterized by highly variability in genetic subtypes associated with IDH1/2 mutations, loss and gains on chromosomes 1, 7, 10, and 19. With combined multi-omic SOM portrayal we showed that methylation and CNV both affect the expression landscape but in an independent way. Moreover, there was a bias toward the contribution of those modalities depending on the LGG genetic group. Finally, we demonstrated

that the stratification of LGG samples according to the expression and methylation subtypes is more informative from the prognostic point compared to the genetic subtypes.

Of course, this study has some worth noting limitations. First, some regulatory or structural data, such as miRNAs (H. Chen et al., 2023), transcription factors (Zacksenhaus et al., 2017), chromatin modifiers (Zhuang et al., 2020), or topologically associating domains (Campbell, 2019) were not included in our analysis though they are implicated in cancers. These factors are known to be implicated in cancers. Another issue is often imbalanced sample sizes in datasets, that can considerably affect the statistical power of results. Finally, the inclusion of many omic layers can inflate the complexity of downstream calculations.

INFERENCES

1. A ml-SOM pipeline was developed to enable integrative analysis of multi-omic data, extracting perturbed gene modules across omic landscapes, and their functional annotation. It also enables the analysis of mutual associations between functional (gene expression) and regulatory modalities (such as methylation) as well as genomic features (including copy number and single nucleotide variations) and maps these features to phenotype and clinical data.
2. A supSOM pipeline was developed to enable a transfer learning approach to project new samples into existing SOM space.
3. Breast cancer PAM50 subtypes show subtype characteristic perturbations of gene modules across expression, methylation, copy number, and single nucleotide variations that are associated with subtype-specific survival and clinical outcomes.
4. Multi-omic SOM portrayal of low-grade gliomas showed better resolution of molecular subtypes and prognostic indicators of omic-based subtypes compared with WHO genetic subtypes.

LIST OF PUBLICATIONS AS A PART OF DISSERTATION TOPIC

1. **Davitavyan S.** Multi-omics portrayal of breast cancers. Вестник РАУ. 2024; N 1:50-57.
2. Kryukov, K.; **Davitavyan, S.**; Stupichev, D.; Sharun, A.; Love, A.; Kleimenov, M.; Kuznetsov, S.; Tkachuk, A.; Kushnarev, V.; Abstract 4922: Unraveling sarcomatoid features in clear cell renal cell carcinoma with RNA-seq. Cancer Res 15 March 2024; 84 (6_Supplement): 4922. https://doi.org/10.1158/1538-7445.AM2024-4922
3. Kushnarev, V.; Stupichev, D.; Kryukov, K.; **Davitavyan, S.**; Johnson, M.; Shanthappa, B.U.; Xiang, Z.; Nomie, K.; Postovalova, E.; Bagaev, A.; Fowler, N.; Meric-Bernstam, F.; 143 Correlating RNA-seq detection and IHC staining of potential antibody-drug conjugate (ADC) targets: HER3, HER2, TROP2, Nectin4, and aFLR. BMJ Specialist Journals 2023; 11(163); https://doi.org/10.1136/jitc-2023-SITC2023.0143
4. Binder, H.; Schmidt, M.; Hopp, L.; **Davitavyan, S.**; Arakelyan, A.; Loeffler-Wirth, H. Integrated Multi-Omics Maps of Lower-Grade Gliomas. Cancers 2022, 14, 2797. https://doi.org/10.3390/cancers14112797
5. Nikoghosyan, M.; Loeffler-Wirth, H.; **Davitavyan, S.**; Binder, H.; Arakelyan, A.; Projection of High-Dimensional Genome-Wide Expression on SOM Transcriptome Landscapes. BioMedInformatics 2022, 2, 62-76. https://doi.org/10.3390/biomedinformatics2010004

ԴԱՎԻԹԱՎՅԱՆ ՍՈՒՐԵՆ ՍԱՄՎԵԼԻ

ԿՐԾՔԱԳԵՂԻ ՔԱՂՑԿԵՂԻ ԵՎ ԳԼԻՈՄԱՅԻ ՄՈԼԵԿՈՒԼԱՅԻՆ ԲԱԶՄԱԶԱՆՈՒԹՅԱՆ ԲՆՈՒԹԱԳՐՈՒՄԸ ՏՐԱՆՍԿՐԻՊՏՈՄԱՅԻՆ, ԳԵՆՈՄԱՅԻՆ ԵՎ ԷՊԻԳԵՆԵՏԻԿԱԿԱՆ ՏՎՅԱԼՆԵՐԻ ՀԻՄԱՆ ՎՐԱ

ԱՄՓՈՓԱԳԻՐ

Հանգուցային բառեր՝ մեքենայական ուսուցում, ինքնակազմակերպվող քարտեզներ, կրծքագեղձի քաղցկեղ, գլիոմա

Վերջին տարիներին ուռուցքների զարգացման մեխանիզմներն ուսումնասիրող բազմաթիվ հետազոտությունների շնորհիվ սկիզբ է դրվել բժշկության մեջ նոր դարաշրջանի: Չնայած այս ուսումնասիրությունների արդյունքում մշակվել են քաղցկեղների բուժման նոր մոտեցումներ՝ բարձրացնելով բուժառուների ապրելիությունը, այսօր էլ հրատապ անհրաժեշտություն կա քաղցկեղների համապարփակ ուսումնասիրման՝ հիվանդության արդյունավետ կանխատեսման, կանխարգելման, ինչպես նաև հիվանդների ապրելիության գնահատման և բուժման նպատակով: Նոր գիտելիքների կուտակմանը զուգընթաց ակնհայտ է դառնում, որ քաղցկեղը դժվար է հետազոտել և բուժել հիվանդության բարդ ու տարասեռ բնույթի պատճառով: Ներկայում քաղցկեղային հետազոտությունները հիմնականում ուղղված են քաղցկեղի տարբեր տեսակների մոլեկուլային բազմազանությունը (միջուռուցքային տարասեռություն) բացահայտելուն՝ բնակչության ավելի լայն զանգվածների համար բուժման մատչելիությունը բարձրացնելու համար: Այդուհանդերձ, ներուռուցքային տարասեռությունը առանցքային նշանակություն ունի անհատականացված բժշկության մեջ և գիտական այս ոլորտում, քանի որ նպաստում է քաղցկեղի տարբեր ենթատեսակներում նոր առանձնահատկությունների բացահայտմանը:

Տրանսկրիպտոմային, գենետիկական և էպիգենետիկական մակարդակներում հայտնաբերվել և նկարագրվել են բազմաթիվ փոփոխություններ տարատեսակ քաղցկեղային հիվանդություններում: Գլիոմաները և կրծքագեղձի քաղցկեղը բացառություն չեն, որոնց բարձր ներուռուցքային տարասեռությունից ենթադրվում է, որ պոտենցիալ արդյունավետ բուժումները կարող են անտեսվել՝ որոշակի մոլեկուլային վարիացիաների աննկատ մնալու պատճառով: Մինչդեռ քաղցկեղի բուժման մոտեցումները ուղղակիորեն փոխկապակցված են մոլեկուլային առանձնահատկությունների հետ: Ուստի խիստ անհրաժեշտություն կա մշակելու այնպիսի գործիքներ, որոնք հաշվի կառնեն այդ առանձնահատկությունները՝ հնարավորություն տալով իրականացնել քաղցկեղի մոլեկուլային բազմազանության համապարփակ վերլուծություն:

Ատենախոսական աշխատանքի նպատակն է բազմաօմիկ տվյալների վրա հիմնված մեքենայական ուսուցման ալգորիթմների մշակումը և կիրառումը՝

կրծքագեղձի քաղցկեղի և գլիոմայի մոլեկուլային բազմազանությունը նկարագրելու համար:

Քաղցկեղային նմուշների փոփոխությունների հիմքում ընկած գործոնների բացահայտումը հիմք է հանդիսանում նպատակային և անհատականացված բուժական մոտեցումներ մշակելու համար՝ նախատեսված քաղցկեղի սպեցիֆիկ ենթատեսակի և ընդհուպ մինչև մեկ անհատի համար: Չնայած տարբեր քաղցկեղների վերլուծության համար մշակվել են մի շարք ալգորիթմներ, դրանք հաճախ սահմանափակվում են ընդհանուր հետազոտությամբ՝ առանց հաշվի առնելու քաղցկեղների տարասեռ բնույթը: Մեր կողմից մշակված մեքենայական ուսուցման մոտեցումը թույլ է տալիս լուծել այս խնդիրը՝ մասնավորապես, բացահայտել քաղցկեղների մոլեկուլային առանձնահատկությունները, կապ հաստատել նմուշներում հայտնաբերված փոփոխված գործոնների և կլինիկական հետազոտությունների ընթացքում գրանցված փոփոխությունների միջև:

Նորամշակ մոդելները գեներացրել են նշանակալի արդյունքներ քաղցկեղի երկու տեսակների համար և վալիդացվել են համանման արդյունքներով այլ հետազոտություններում, ինչը վկայում է մեթոդի արժանահավատության մասին:

Մեթոդի կիրառման արդյունքում բացահայտվել է ասոցիացիաների սպեցիֆիկություն կրծքագեղձի քաղցկեղի տարբեր ենթատեսակներում, ինչպես նաև առանձնացվել են որոշակի գեների կլաստերներ և բացահայտվել դրանց ասոցիացիան ինչպես ենթատեսակների, այնպես էլ կլինիկական պարամետրերի և ապրելիության հավանականության հետ: Կրծքագեղձի քաղցկեղի դեպքում գեների էքսպրեսիայի փոփոխությունները հիմնականում նույնաուղղված են, և տարբերությունները ավելի շատ քանակական են, քան՝ որակական, իսկ մեթիլավորումը, էքսպրեսիայի համեմատ, առավել մեծ տատանում է ցուցաբերում տարբեր ենթատեսակներում: Պատճենների քանակի փոփոխությունը ընդհանուր բնույթ է կրում բոլոր տիպերում, մինչդեռ եզակի նուկլեոտիդային պոլիմորֆիզմների (SNP) փոփոխությունը աննշան է: Գլիոմաների համապարփակ բազմաօմիկ հետազոտության արդյունքում արձանագրվել է, որ նմուշների դասակարգման համար առավել նպատակահարմար է օգտագործել տրանսկրիպտոմային և էպիգենետիկական դասակարգումները, քան գենետիկական: Մշակված հավելյալ մեթոդը (sup-SOM) հաջողությամբ կարող է ներառել նոր նմուշներ արդեն «ուսումնառված» մոդելում, ինչը թույլ է տալիս շրջանցել վերջինիս կրկնակի «ուսումնառությունը»:

Այսպիսով, իրականացված աշխատանքը կարող է բարելավել քաղցկեղային հիվանդությունների դասակարգման մեթոդները, ինչպես նաև հիմք հանդիսանալ անհատականացված բժշկության մոտեցումների մշակման համար:

ДАВИТАВЯН СУРЕН САМВЕЛОВИЧ

ХАРАКТЕРИСТИКА МОЛЕКУЛЯРНОГО РАЗНООБРАЗИЯ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ И ГЛИОМЫ НА ОСНОВЕ ТРАНСКРИПТОМНЫХ, ГЕНОМНЫХ И ЭПИГЕНЕТИЧЕСКИХ ДАННЫХ

РЕЗЮМЕ

Ключевые слова: машинное обучение, самоорганизующиеся карты, опухоль молочной железы, глиома

В последние годы многочисленные исследования, изучающие механизмы развития опухолей, привели к возникновению новой эры в медицине. Несмотря на то, что эти исследования разрабатывают новые подходы к лечению и повышают выживаемость больных, существует острая необходимость в всестороннем изучении болезни, что в свою очередь позволит успешно прогнозировать и предотвращать заболевания, а также лечить пациентов. По мере накопления новых знаний становится ясно, что опухоль достаточно трудно исследовать и лечить из-за его сложности и неоднородности. Исследования в основном направлены на выявление молекулярного разнообразия (межопухолевой гетерогенности) различных типов опухоли, чтобы повысить доступность лечения для более широких слоев населения. С другой стороны, внутриопухолевая гетерогенность имеет ключевое значение в области персонализированной медицины и научных исследований, поскольку она способствует выявлению новых особенностей в различных подтипах.

Было обнаружено и описано множество изменений на транскриптомном, генетическом и эпигенетическом уровнях при различных опухолевых заболеваниях. Глиомы и опухоли молочной железы не являются исключением, из-за высокой внутриопухолевой неоднородности этих видов рака предполагается, что потенциально эффективными методами лечения можно пренебречь из-за того, что определенные молекулярные изменения могут остаться незамеченными. Подходы к лечению коррелируют с молекулярными особенностями, поэтому существует острая необходимость в разработке таких инструментов, которые могут учитывать особенности и проводить всесторонний анализ молекулярного разнообразия заболевания. Такие инструменты уже широко используются в научной и клинической областях. Однако инструменты биоинформатики в основном предназначены для выявления особенностей только на определенных уровнях. С другой стороны, подходы, основанные на машинном обучении, в основном решают проблему классификации пациентов или прогнозирования типа заболевания у пациентов с неизвестными диагнозами.

Целью исследования является разработка алгоритмов машинного обучения, основанных на мультиомных данных, и их применение для описания молекулярного разнообразия опухоли молочной железы и глиомы.

Выявление факторов, лежащих в основе изменений в опухолевых образцах, является основой для целенаправленного и персонализированного лечения, предназначенного для определенных подтипов и даже для одного человека. Несмотря на то, что было разработано множество моделей для анализа различных видов опухоли, они часто ограничиваются общими исследованиями из-за неоднородности и сложности заболеваний. Разработанный нами подход машинного обучения позволяет решить эти проблемы. В частности, этот метод предназначен для выявления молекулярных особенностей рака, устанавливает связь между

измененными факторами, обнаруженными в образцах, и изменениями, зарегистрированными во время клинических исследований.

Разработанные модели показали значимые результаты для обоих типов опухоли, которые были подтверждены аналогичными результатами в других исследованиях, что говорит о достоверности подхода. При использовании метода были выявлены специфичные ассоциации в различных подтипах опухоли молочной железы, а также выделены кластеры определенных генов и была идентифицирована их связь как с подтипами, так и с клиническими параметрами и выживаемостью. В случае опухоли молочной железы изменения в экспрессии генов в основном однонаправлены, и различия скорее количественные, чем качественные, а метилирование по сравнению с экспрессией, проявляет наибольшую вариабельность в разных подвидах. Изменение количества копий генов носит общий характер для всех подтипов, в то время как изменения однонуклеотидных полиморфизмов (SNP) незначительны. В результате комплексного исследования глиом было зафиксировано, что классификация образцов по значениям экспрессии и метилирования является более информативной, чем генетическая классификация. Разработанный дополнительный метод (sup-SOM) может успешно интегрировать новые образцы в обученную модель, что позволяет обойти двойное обучение модели.

Таким образом, проведенная работа имеет потенциал для улучшения методов классификации, а также может стать основой для улучшения подходов персонализированной медицины.