Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Unsupervised breast cancer subclass discovery from aberrant glycogene expression for reliable clinical prognosis

Mia Venter
University of Cape Town
Kevin Naidoo
Kevin.Naidoo@uct.ac.za

University of Cape Town

Research Article

Keywords:

Posted Date: July 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10155568/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Unsupervised breast cancer subclass discovery from aberrant glycogene expression for reliable clinical prognosis

Mia Venter[1,2] and Kevin J. Naidoo[1,2,*]

[1]Scientific Computing Research Unit Address, PD Hahn Building, University of Cape Town, Rondebosch 7701, South Africa [2]Department of Chemistry, PD Hahn Building, University of Cape Town, Rondebosch 7701, South Africa.

*Corresponding author Kevin J. Naidoo e-mail: kevin.naidoo@uct.ac.za

Abstract

Unsupervised hierarchical clustering of breast cancer glycogene expression data using the Growing Hierarchical Self Organising Representation Map algorithm identified six subclasses. These glycogene expression clusters (GEC) overlap with traditional markers such as PAM50 subtypes and the mRNA stemness index but provide improved prognostic stratification. Feature discovery via Recursive Feature Elimination with Strikeout identified glycogenes that are important predictors of each GEC. The results reveal that distinct glycogene signatures exist within each GEC, which are associated with unique oncogenic functionalities. These glycogene signatures covary with immune cell type abundance, which may point to context-specific interplay between immunity and glycosylation. The functional aberrations in each GEC may be targeted by specific drugs, offering a proof-of-concept for functionally informed, personalised drug discovery. This study demonstrates how using glycogene expression in combination with specific machine learning algorithms could be used as an approach for more personalised cancer subtyping.

Introduction

Gene-expression based cancer subtyping inherently suffers from the curse of dimensionality, where the sample size is several orders of magnitude smaller than the number of genes (1). Artificial intelligence (AI) and machine learning (ML) algorithms are increasingly being used to

radically reduce the dimensions of gene sets and cluster the samples into clinically meaningful classes.

As machine learning methods have advanced, undertaking breast cancer subtyping on the basis of gene expression has become a reliable approach (2). Several gene expression procedures are utilised in clinical settings for breast cancer subtype diagnostics, including the PAM50 intrinsic molecular subtyping technique (3, 4). The PAM50 classes are well established and have been used in an analysis of the stratification of cancer samples based on the degree of differentiation, otherwise known as the mRNA stemness index (mRNAsi) (5). Despite these advances, a biological description of patient prognosis and an understanding of the treatment regimen that responds to the underlying biochemical fundamentals driving tumour progression, remains unmet (6).

Glycosylation, a non-template driven, highly orchestrated post-translational modification, is dysregulated in cancer (7, 8). The glycoconjugates that present themselves on the surface of epithelial cells as signalling (9) and immune receptors (10), play various roles in tumorigenesis (7, 11-13). The glycogenes that catalyse these reactions are located in specific areas along the ER-Golgi complex, and their abundance and spatiotemporal location determine the nature of protein and lipid glycosylation (14-16). Glycogene expression can classify different cancer types and cluster breast cancer samples into clinically relevant groups (17). The pervasiveness of glycosylation in cellular functionality and its correlation to hormonal and developmental aberrations provides an opportunity for improved breast cancer subtyping. Because tumour glycosylation patterns closely align with established classifiers like PAM50 and stemness (18), a glycosylation-based subtyping method could effectively complement current diagnostic tools.

The Growing Hierarchical Self-Organising Map (GHSORM) (19) was developed to identify hierarchical class-representative structures in datasets such as breast cancer gene expression in this study. The algorithm first performs dimensionality reduction using a Denoising Autoencoder that ensures representative learning from high-dimensional data. Following this, the representation layer is mapped onto a series of hierarchically structured Self-Organising Maps (SOMs), each layer offering finer-grained resolution of the clustering space than the preceding

one. GHSORM outperforms more rudimentary clustering techniques such as k-means and hierarchical clustering, particularly in cases where the data is hierarchical (19).

A rational explanation of ML-derived subtype models is essential for informing patient care following diagnoses (20). While unsupervised methods such as GHSORM are useful to uncover underlying data structures, they lack information regarding important genes or possible disease-causing mechanisms. We combine our unsupervised clustering algorithm with an in-house feature discovery technique based on the Random Forest algorithm (21), Recursive Feature Elimination with Strikeout (RFES), to provide the rationale for clustering. Building on this, biological knowledge and bioinformatics are applied to identify potential mechanisms underlying the biology of disease clusters, as well as new targets for functional validation.

Results

Glycogene expression-based breast cancer subtype discovery with GHSORM

Six Glycogene Expression Clusters (GECs) were identified following the application of the unsupervised GHSORM subtype discovery algorithm to the TCGA data **(Figure 1A)**. The first SOM layer immediately separates Basal from Luminal cancers, while the second layer offers a more granular clustering scheme. The ideal number of clusters was determined through assessing the hierarchical distance between neurons, analysing the homogeneity of glycogene expression across clusters, and maximising patient survival stratification between GECs **(Supplementary Figure S1)**.

There are currently no existing standards for the hyperparameterisation of unsupervised models, yet the process is highly subjective and has a significant impact on model performance (18). Due to the clinical relevance of the PAM50 subtypes, and the observed overlap between PAM50 subtype and glycogene expression (17), GHSORM was hyperparameterised by maximising the Normalised Mutual Information (NMI) index (22) between GECs and PAM50 subtype (See Experimental Section). Through this approach, the best-performing GHSORM model had a Normalised Mutual Information (NMI) index of 0.63. In contrast, the application of static clustering techniques such as DASOM (23) and k-means clustering resulted in inferior model performance (NMI = 0.41 and 0.23 respectively) **(Supplementary Figure S2)**.

Due to the established connection between glycosylation and stemness (18), we investigated whether the GECs stratified samples based on mRNAsi (5) **(Figure 1B; Supplementary Table S1)**. The interaction between GEC and PAM50-mRNAsi was significant for the LumA, LumB, and Basal subtypes (chi-square p = 2.187e-07, 4.812e-07, and 4e-04, respectively), but was less significant in the case of the HER2 subtype (chi-square p = 5e-02) and insignificant for the Normal-like subtype as only one sample was high-stem (mRNAsi 0.5). Therefore, the distinction between high and low mRNAsi was not further investigated for the HER2 and Normal-like subtypes.

A key metric of effective subtyping is the ability to stratify patient survival; in this regard, the glycosylation-based approach outperformed both PAM50 and the PAM50-mRNAsi combination **(Figure 1C; Supplementary Figure S3)**. However, none of the approaches reached statistical significance ($p > 0.05$). This is due to the exclusion of stage III and IV samples from the dataset. When stage III and IV samples are included, glycosylation-based clustering results in statistically significant patient stratification ($p = 0.01$), while PAM50 and the combination of PAM50-mRNAsi are less significant ($p = 0.1$ and 0.05, respectively) **(Supplementary Figure S4)**. The decision to remove stage III and IV samples from the cohort was made to retain emphasis on the utility of glycosylation-based subtyping in the early stages of breast cancer.

The existence of these GECs was validated in the METABRIC dataset by following the same clustering procedure as for the TCGA dataset (See Experimental Section). By using the Jaccard index to measure the similarity of the differential expression signature between the two datasets, 5 of the 6 GECs could be found in METABRIC and corresponded with those discovered in the TCGA. The reason for GEC6 missing in the METABRIC dataset was likely due to the small number of Basal samples (n = 40) **(Supplementary Figure S5)**.

Recursive Feature Elimination with Strikeout (RFES)

To determine which glycogenes significantly contribute to GEC classification, RFES was developed. RFES iteratively removes unimportant genes from a Random Forest classification model, terminating when model accuracy decreases *s* times. The introduction of the *strikeout* parameter allows the accuracy to fluctuate, thereby increasing the flexibility of the model and facilitating a more exhaustive search of the feature space.

Using RFES, 60 glycogenes were identified that could classify samples into GECs with 95.45% accuracy using a 5-fold cross-validation approach **(Supplementary Figure S6)**. The glycosignature, defined as a subset from the 353 total GEGs, differs notably for each of the six GECs **(Figure 2A)**, and the top 20 glycogenes important for the classification of each GEC are distinctly different **(Figure 2B)**. Moreover, model performance remains consistent between classes **(Figure 2C)**.

The immune landscape of GECs

Undertaking immune cell deconvolution using Cibersort revealed extensive variation of overall immune infiltration across GECs, as indicated by Cibersort's absolute score **(Figure 3A)**. Expectedly, overall immune infiltration and tumour purity show a strong negative correlation across the GECs **(Supplementary Figure S7)**, and tumour purity was strongly associated with GEC (Anova p-value < 2e-16). Therefore, to quantify observed differences in relative immune cell abundance **(Figure 3B)**, CLR-normalised relative immune cell abundance was modelled as a function of GEC, adjusting for tumour purity. Five immune cell types were significantly different in one GEC compared to other GECs combined (adjusted $p < 0.05$). GEC4 was characterised by a low abundance of M0 macrophages, while GEC1 had a low abundance of M1 macrophages. M2 macrophages were most abundant in GEC2, and resting mast cells were present at low abundance in GEC6 **(Figure 3C)**.

The strong regulation of immune cells by glycosylation likely underpins the ability of glycogene expression to produce breast cancer subtypes that vary in terms of immune infiltration, tumour purity, and estimated abundance of immune cells. Thus, the covariation of glycogene paths and relative immune cell abundance was investigated through Pearson correlation **(Figure 3D)**. Patterns of covariation differed by GEC. M0 macrophage abundance in GEC4 negatively correlated with N- and O-linked glycan degradation, a pattern not observed as strongly in other GECs. Furthermore, resting mast cell abundance in the Basal clusters GEC5 and GEC6 exhibited opposing correlations with 6-sulfated Sialyl Lewis X formation – positive in GEC5 and negative in GEC6.

Gene regulatory network and drug discovery

The heterogeneity observed in breast cancer is indicative of a variety of tumorigenic mechanisms, each likely accompanied by distinct glycosylation aberrations. To unpack this, a gene regulatory network for breast cancer was constructed where regression modelling was used to predict the expression of non-glycogenes from RFES-identified and differentially expressed glycogene targets in each GEC (See Experimental Section). The ability of a glycogene to accurately predict the expression of a non-glycogene may indicate a functional relationship and presents a target for future research. Functional profiling of the non-glycogenes was conducted to reveal potential associations for glycosylation in each GEC.

Different glycogenes are important predictors of gene expression in each GEC **(Figure 5A)**, and those genes have distinct functional roles **(Figure 5B)**. In GEC1, the expression of glycosaminoglycan biosynthesis genes such as CHST11 could predict the expression of immune-related genes such as the T-cell surface glycoprotein CD5, SIGLEC7, and CD80. In GEC3, FUT8 could predict the expression of PEX12, a peroxisomal biogenesis factor. Glycogene expression and relation to functional processes varies across GECs, suggesting that glycosylation may play various roles in the tumorigenic process.

The major value of subtype discovery is to provide a clinically meaningful direction which results in a treatment rationale that is unique and effective for each subtype. To identify potential treatments for the GECs, Functional Representation of Gene Signatures (FRoGS) was utilised (24). FRoGS works by predicting compound-target interactions based on the concordance between gene signatures in a sample and the signatures a compound is known to inhibit. Cancer drugs that show potential for repurposing were modelled for their ability to inhibit the functional gene signatures predicted from glycogene expression in each GEC **(Figure 5C)**. From this modelling, Albendazole was predicted to be an effective treatment for GEC2, while Clomifene may be effective for patients belonging to GEC4.

Discussion

Luminal and Basal cancers are known to be genetically distinct (25), and aberrant glycogene signatures drive their separation in the first layer SOM **(Figure 1A)**. GHSORM's probing of the

more subtle remaining differences gave rise to two Basal and two Luminal clusters which emerged in the organisational hierarchy of the second layer SOM. Although the GECs overlap with traditional classification by PAM50 and mRNAsi **(Figure 1B; Table 1)**, the superior and custom-designed clustering algorithm applied to glycogenes enables improved patient survival stratification compared to these traditional approaches **(Figure 1C, Supplementary Figure S3 and S4)**. The existence of comparable GECs in the METABRIC dataset show that these glycogene signatures are stable and reproducible **(Supplementary Figure S4)**, motivating further research into glycogene-based cancer subtyping.

The concordance between clusters in the METABRIC and TCGA datasets might be improved if the same set of glycogenes were used to cluster both datasets **(Table 1)**. Many glycogenes critical to the clustering of the TCGA dataset, such as B3GNT5, CHST11, and ST6GAL2 were not present in the METABRIC dataset. We made this experimental design choice to maximise the information gain from the TCGA dataset, at the expense of concordance with the METABRIC cohort, which contained fewer glycogenes.

One major barrier to the adoption of machine learning in clinical research is lack of model explainability (26). RFES was therefore developed to identify glycogenes that contribute to GHSORM's clustering scheme **(Supplementary Figure S6)**. Key considerations for RFES performance were the maximisation of the model accuracy on an independent testing dataset, minimisation of the per-class classification error, and minimisation of the number of genes in the final model. Each GEC could be classified by a distinct set of glycogenes **(Figure 2)**, providing opportunities for glycogene-based breast cancer subtyping. Although beyond the scope of this paper, follow-up mechanistic studies that investigate how these glycogenes are involved in tumorigenesis in each subtype is critical to establish the feasibility of glycosylation-based subtyping in clinical settings.

An important objective of this paper was to investigate the immune phenotypes associated with each GEC. This is of particular interest given the tight integration between glycosylation and the immune system (27). Unsurprisingly, tumour purity negatively correlated with immune infiltration across the GECs **(Supplementary Figure S7)**. After adjusting for tumour purity, the

degree of macrophage differentiation was different between GECs. Solid tumours are known to be heavily infiltrated by macrophages, but their differentiation state depends on the tumour microenvironment (28). Compared to the other GECs, GEC2 has a high abundance of M2 macrophages **(Figure 3C)**, which are typically associated with adverse patient outcomes due to their ability to promote immunosuppression and angiogenesis (29).

Interestingly, covariation patterns between glycosylation pathways and immune cell abundance varied between the GECs **(Figure 3D)**. This finding provides mechanistic clues about the context-specific interplay between glycosylation and immune system regulation that could be further investigated through functional validation. For example, the biosynthesis of 6-sulfated Sialyl Lewis X structures showed a positive correlation with resting mast cells in GEC5, but a negative correlation in GEC6, both Basal clusters. Mast cell activity is controlled via binding of Siglec-8 to 6-sulfated Sialyl Lewis X (30). This finding illustrates how glycosylation-based subtyping could provide more information than PAM50 subtype alone, which is especially important given the urgency of improved care for triple negative breast cancers (31).

In accordance with the proxy-hypothesis (32), we hypothesised that covarying expression patterns between glycogenes and non-glycogenes represent functional networks that drive tumorigenesis in a coordinated fashion. Functional profiling of the constructed gene regulatory network showed distinct covarying patterns of glycogene expression between GECs **(Figure 4A, B)**. In GEC1, for example, expression of glycogenes involved in glycosaminoglycan synthesis and degradation such as CHST11 predicted the expression of key immune modulators such as CD5, CD80 and SIGLEC7. Although the exact mechanistic roles for glycosaminoglycans in immune regulation is still unclear, significant effort is underway to disentangle these complex networks (33).

Through a deep learning approach that predicts cancer treatment based on the functional profile of cancer samples (24), potential drugs were identified for each GEC **(Figure 4C)**. Based on the model, Clomifene may be an effective drug against GEC4. Clomifene showed anti-proliferative capacity and induced apoptosis in cancer cells (34). While further studies are needed to validate these compounds as anti-cancer agents specific to each GEC, this analysis

provides a proof-of-concept showing how the integration of expression patterns and functional profiling can be leveraged to discover potential drug candidates.

This study has limitations that must be acknowledged. The relationship between glycogene expression and glycosylation is not yet well-established; therefore, we refrain from inferring the glycomic state of the GECs. In addition, the samples used in this study are derived from bulk tumour tissue and thus the specific cell types that contribute to the observed glycosylation signatures are not known. Functional validation studies linking glycosylation to known tumorigenic pathways are critical for translation of this work from theory to practice.

This research serves as a proof-of-concept, demonstrating that specialised machine learning algorithms can be paired with aberrant glycogene expression to discover distinct cancer subtypes. By linking molecular expression profiles with patient prognosis and tailored treatment strategies, this approach establishes a robust framework for data-driven cancer subtype discovery.

Materials and Methods

Data preparation

The TCGAbiolinks package (35) in R-4.4.1 (36) was used to download 1067 primary, untreated breast cancer and 111 matched normal RNA-seq samples from female patients from The Cancer Genome Atlas (https://www.cancer.gov/tcga). Ensemble IDs were converted to HGNC symbols using the org.Hs.eg.db library (37), and in cases where duplicate symbols exist, the entry with the highest mean expression samples was used. GE genes and pathways were obtained from GlyocoEnzOnto (38). Tidyverse (39) and data.table (40) were instrumental in performing dataset manipulation.

The mRNA stemness index (5) (mRNAsi) was obtained using the “TCGAanalyze_Stemness” function and SC_PCBC_stemSig score from TCGAbiolinks (35). Samples were designated as Low Stem (LS; mRNAsi < 0.5) or High Stem (HS; mRNAsi >=0.5).

Differential expression analysis was conducted using the DESeq2 (41) package and Log Fold changes were obtained using the "apeglm" method (42). The variance stabilised data (using only the intercept in the design matrix) was used for further analysis.

Sample selection and experimental setup

The Cancer Genome Atlas (TCGA) breast cancer dataset contains 1067 female treatment-naïve primary tumour RNA-seq samples. To ensure the applicability of our method to the stratification of early breast tumours, only 785 Stage I and II samples were used in this analysis (n_{total} =785; n_{LumA} =406; n_{LumB} =139; n_{Her2} =57; n_{Basal} =155; $n_{Normal\text{-}like}$ =28).

Evaluating the performance of unsupervised clustering models is challenging due to the lack of a ground truth. Since previous research demonstrated overlap between glycogene expression and PAM50 subtype, GHSORM performance was evaluated by maximising the Normalised Mutual Information (NMI) index between Glycogene Expression Clusters (GECs) and PAM50 subtype.

To mitigate against the class bias caused by LumA overrepresentation, only 147 LumA samples were included in the final dataset. This number of LumA samples were chosen as the average of the two next-biggest subtype classes, Basal and LumB. To ensure an even distribution of LumA samples in terms of glycogene expression, GHSORM was first run on the LumA samples only (n_{LumA} =406) to determine the number of LumA glycogene clusters **(Supplementary Figure S8)**. LumA samples in the final dataset were randomly and proportionally sampled from the 3 identified LumA glycogene clusters, yielding the final dataset (n_{total} =526; n_{LumA} =147; n_{LumB} =139; n_{Her2} =57; n_{Basal} =155; $n_{Normal\text{-}like}$ =28).

The METABRIC (43) dataset was used for validation of the GECs. To make the analysis comparable, only stage I and II primary tumours from female patients were included. PAM50 subtype information was obtained from CBioPortal (44, 45), and for samples that lacked a PAM50 subtype tag, as well as for those designated claudin-low, the genefu algorithm was used(46). A similar approach was followed to down-sample the LumA subtype **(Supplementary Figure S8)**, yielding the final METABRIC dataset (n_{total} =189; n_{LumA} =43; n_{LumB} =46; n_{Her2} =29; n_{Basal} =40; $n_{Normal\text{-}like}$ =31).

The Growing Hierarchical Self-Organising Map

We developed the Growing Hierarchical Self-Organising Representation Map (GHSORM) specifically for the purpose of cancer subtype discovery (19). GHSORM relies on two components: the Denoising Autoencoder (DA) and the Self-Organising Map (SOM). These components can be combined in different ways to create several algorithmically similar models. To find the optimal algorithmic structure and model hyperparameters, a brute-force grid search approach was employed with the objective of maximising the Normalised Mutual Information (NMI) between clusters, with PAM50 subtype as the class label. The best-performing model was a Pre-Processor GHSORM in which one DA is shared between all SOM layers. All glycogenes with a standard deviation (SD) > 0.5 across samples was used as input to GHSORM.

At the encoder end, glycogene expression input features were corrupted by addition of Gaussian noise with a standard deviation of 0.2. The contribution of the corrupted input feature to the hidden layer was neuron and is defined by weights, W and biasing parameters, b. A hyperbolic tangent function, f, is applied to these that give the activation of the hidden neurons.

$$y = f(W\tilde{x} + b)$$

(1)

At the decoder end (the output layer) the samples' reconstructed features, $\hat{x}$, are generated. This is done via a set of parameters biasing the neurons, $\hat{b}$, in the network's layers accompanying the weights, $\widehat{W}$. A hyperbolic tangent function, f, is once again applied to regulate of the contribution of the representation layer, *y*.

$$\hat{x} = f(\hat{W}y + \hat{b})$$

(2)

This procedure of learning features from artificially corrupted input data enables dimensionality reduction while still preserving the structure of the input data space.

The hidden representation layer is then mapped onto a SOM. Training the SOM involves an ordering and a convergence phase, which were implemented using one and ten epochs of samples respectively, according to the following cost function:

$$K(W, b, U) = \min_{d} \sum_{e=1}^{E} h_{de} \, || \, y - \, y_e \, ||^2$$

(3)

Where *E* is the total number of SOM neurons and u_e represents their codebook vectors. The neighbourhood function, h_{de}, controls the surface elasticity of the neuron lattice.

Self-Organising Map growth and depth in the GHSORM algorithm are controlled by the user-provided parameters τ_1 and τ_2 according to the following equations:

$$mqe^{(\sigma)} < \tau_1 qe_{\xi}{}^{(\sigma)_{parent}}$$

(4)

$$qe_e{}^{(\sigma)} < \tau_2 qe_1{}^{(0)}$$

(5)

Where *mqe* represents mean quantisation error, *qe* is quantisation error, and ξ is the index of the neuron of the parent map σ. This resulted in a GHSORM model that promoted architectural depth over the size of individual SOMs in the model. The Pre-Processor GHSORM, wherein all SOMs in the model share a DA, was used because this resulted in maximal Normalised Mutual Information between clusters. GHSORM parameters used in this study can be found in **Table 1**. Pseudocode describing the GHSORM algorithm can be found as **Supplementary Text 1**.

The aweSOM (47) and Kohonen (48) R-packages were used to visualise GHSORM output. Hierarchical superclass clustering of the SOM codebook vectors was performed to determine which SOM neurons belonged in a GEC.

The optimal number of superclasses was determined through visually assessing the hierarchical distance between neurons, analysing the homogeneity of glycogene expression across clusters, and maximising patient survival stratification between GECs.

Validation of GECs using the METABRIC dataset

To compare the GECs identified in TCGA and METABRIC, differential gene expression was performed as for each GEC against the other GECs combined. Glycogenes were extracted, and their regulation was determined (Log2FoldChange > 0 and p-adjusted < 0.1 = upregulated; Log2FoldChange < 0 and p-adjusted < 0.1 = downregulated). The Jaccard index between GECs was computed based on whether the glycogenes were regulated in the same direction in both datasets.

Recursive Feature Elimination with Strikeout (RFES)

To identify the minimum set of glycogenes that can classify the GECs with high accuracy, a Recursive Feature Elimination with Strikeout (RFES) feature discovery wrapper was designed around the Random Forest algorithm. The process commenced with the complete set of 353 glycogenes. In every iteration, unimportant glycogenes were eliminated until model accuracy decreased *s* times. This *strikeout* parameter, inspired by the tuning step in the RandomForestSRC framework (49), introduces flexibility into the model by providing a tolerance in the accuracy search such that a decrease in improvement is allowed *s* times before termination. This allows the accuracy to fluctuate prior to locating the ideal gene set **(Supplementary Figure S9)**.

RFES was run using the RandomForest algorithm (50) in the caret (51, 52) package. The model was trained on 80% of the data using a 5-fold cross validation approach and tested on the remaining 20%. Values of *ntree* included 300, 500, and 1000, and an *mtry* from 1 to 40 was used. Computational time was measured with the tictoc R-package (53).

To avoid a decline in Random Forest classification accuracy due to class imbalanced data, majority classes (n > 88) were randomly undersampled and minority classes (n < 88) were artificially inflated with RACOG (54). RFES with this sampling approach was benchmarked against three commonly used feature discovery algorithms: Recursive Feature Elimination (RFE),

Simulated Annealing (SA), and Genetic Algorithms (GA) (52). RFES performed the best, although it ran significantly slower than the other algorithms. This is in part because the RFES algorithm repeats feature discovery three times to avoid inherent algorithmic randomness and produce stable feature sets **(Supplementary Table S2).**

The RFES algorithm proceeds as follows:

s and $Accuracy_{max}$ are initialised to 0.

The feature importance for each gene across all GECs in $data_{train}$ is obtained using the varImp function from the caret package(51, 52), and the maximum variable importance, as Mean Decrease in Accuracy (MDA) for each gene (*j*) across the GECs *(m)* is recorded.

$$MDA_{j,max} = \max\left(MDA_{j,1}, MDA_{j,2}, \dots, MDA_{j,m}\right) \quad (6)$$

The first quantile (Q_1) is found from the maximum VI across all features, where *p* is the total number of features.

$$Q_1 = Quantile_{0.25}(MDA_{1,max}, MDA_{2,max}, \dots, MDA_{p,max}) \quad (7)$$

The features (*F*) are retained if they have a maximum MDA greater than the first quantile.

$$F = (j \mid MDA_{j,max} > Q_1) \quad (8)$$

Using this feature set *F,* the model is run again and the accuracy of the model on $data_{train}$ and $data_{test}$ are recorded as follows:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (9)$$

If $Accuracy_{test}$ (model accuracy on $data_{test}$) is greater than $Accuracy_{max}$ then $Accuracy_{max}$ = $Accuracy_{test}$, else *s* is incremented by 1. This is repeated until *s* > *strikeout*.

Repeat the procedure *rpt* times and record the most accurate model for each repeat.

An important note is that $data_{test}$ does not include artificial samples from the RACOG, such that the accuracy of the model is only evaluated on real samples.

Immune cell deconvolution

The online version of Cibersort (55) was used in relative mode for immune cell deconvolution, utilising the LM22 signature matrix file and the TMM-normalised counts as input. Batch correction was enabled, quantile normalisation was disabled, and 1000 permutations were used for significance analysis. Only the 252 samples with a p-value < 0.05 were utilised in the analysis. To determine immune abundance differences between GECs, compositional CIBERSORT data was CLR normalised (56) and modelled using a linear model, adjusting for tumour purity.

ssGSEA (57) was employed with the GSVA package (58) to calculate glycosylation pathway scores for the investigation of correlations between glycosylation pathways and estimated relative immune cell abundance in each GEC. Only cells with significantly different abundances according to the linear model (p-adjusted < 0.05) were included, and only correlations coefficients > 0.4 are depicted in the figure. Additionally, redundant glycopaths were removed from the figure. Refer to **Supplementary Table S3** for the correlation matrix showing all glycopaths.

Gene regulatory network

Arboreto (59) was used for the construction of a gene regulatory network, using the vst-normalised count matrix as input. The 60 GE genes identified by RFES as important were used as the target genes **(Supplementary Table S4)**, and the expression of the remainder of the genes

with a SD > 0.5 was predicted. Arboreto produces a list of weights representing the importance of a target gene in predicting the expression of a non-target gene. High weights are potentially indicative of a functional relationship between the two genes.

For each GEC, truly upregulated – as compared to normal tissue and to all other clusters (p-adj < 0.05; LFC > 0) – GEs (RFES importance value > 50) and non-target genes and their weights were extracted from the Arboreto output. The non-target genes with a weight > 5 were subjected to over-representation analysis with ClusterProfiler (60, 61) using all the input genes as the background. Pathways with a p-value < 0.05 were considered significant and the combined functional analysis across all GECs was used as input to Cytoscape (62).

Functional Representation of Gene Signature (FRoGS)

FRoGS (24) was implemented to perform drug target prediction. To train gene embeddings and generate signature embeddings for each GEC, the ClusterProfiler (60, 61) results from the Arboreto (59) functional analysis was used as input. Those embeddings were then used to predict compound targets from the L1000 gene signatures for each GEC, using 60 epochs. Only the MCF7 breast cancer cell line was used. The ReDO_DB database (63) was used to obtain a list of compounds that have shown potential as repurposed cancer treatments, and only those compounds were used for further analysis. As output, FRoGS generates files containing a list of target genes ranked by their predicted scores for each compound. For each GEC, compounds were ranked based on the mean predicted scores of the genes, and compounds with a mean score > 0.96 were included in the visualisation.

Other software packages

Survival was analysed using and the survival (64), GGally (65), and survminer (66) R packages. Heatmaps were drawn using ComplexHeatmap (67). Dendsort (68) and circlize (69) were used for hierarchical clustering of heatmap rows and columns, and generation of colour vectors respectively. Circlize (69) was also used to construct the chord diagram. Rstatix was used for statistical testing(70). The pROC package was used to construct Receiver Operating Characteristic curves curves (71). Functional analysis and pathway enrichment was performed using the Gene Ontology database(72, 73), except in the case of glycosylation pathways, which

were obtained from GlycoEnzOnto (38). The Reactome (74, 75) and Genecards (76, 77) databases were frequently used to look up gene function. Figures were compiled with BioRender.com

Funding sources and disclosure of conflicts of interest

This work is based in part upon research supported by the National Research Foundation (NRF) CPPR 466624 grant (KJN) and the South African Medical Research Council Self-Initiated Grant (SAMRC SIG) 416090 grant. MV thanks the National Research Foundation (NRF) for a Doctoral Postgraduate Scholarship (PMDS22051711715), as well as the University of Cape Town for a Doctoral Research Scholarship. None of the authors have competing interests to declare.

Acknowledgements

We thank the Centre for High Performance Computing (CHPC) for computational resources (CHEM0840), and Christos Ferles for assistance with the application of the GHSORM algorithm.

Abbreviations

GEC – Glycogene expression cluster

GHSORM – Growing Hierarchical Self-Organising Representation Map

mRNAsi – mRNA stemness index

NMI – Normalised Mutual Information

RFES – Recursive Feature Elimination with Strikeout

Ethics Consent to Participate, and Consent to Publish declarations:

Not applicable.

Data Availability Statement

The results shown here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga and the Molecular Taxonomy of Breast Cancer International Consortium(43). All R code used in this study will be made available upon request.

Author contributions

This research project benefited from the collaborative efforts of the authors, with distinct contributions as follows: Conceptualisation: KJN. Methodology: MV, KJN. Data curation and analysis: MV. Writing—original draft: MV: Final draft: MV, KJN. Supervision: KJN. Project administration: KJN. Funding acquisition: KJN.

Figure legends

Figure 1: Unsupervised glycogene-based subtype discovery. (A) Glycogene Expression Clusters (GECs) identified by GSHORM. Pie charts represent distribution of PAM50 subtypes and mRNAsi in each Self-Organising Map (SOM) node, and samples that cluster in a single node share similar glycogene expression signatures. Clusters are denoted by background colours. **(B)** Proportion of PAM50 subtype and mRNAsi in each cluster. **(C)** Kaplan Meier plot of patient survival between the GECs. Significance is analysed with the Log rank test.

Figure 2: Supervised feature discovery with RFES. (A) Expression of the 60 glycogenes from the best-performing RFES model. **(B)** Variable importance for the top 20 most important genes in each GEC. **(C)** Receiver Operating Characteristics curve showing performance of the classifier for each cluster.

Figure 3: Immune infiltration and cell abundance across GECs. (A) Cibersort-estimated absolute immune score (Kruskal-Wallis $p < 6.8e\text{-}11$). **(B)** Barplot of average estimated cell proportions in each GEC. **(C)** Relative abundance of immune cell types in one GEC compared to the other GECs combined. Only significant immune cell type-GEC combinations are shown (p-adjusted < 0.05 after adjusting for tumour purity). **(D)** Pearson correlation coefficients between glycosylation pathways and estimated immune cell abundance for significant immune cell types as identified in **(C)**.

Figure 4: Association of differentially glycogenes with biological pathways and drugs candidates. (A) Heatmap showing the importance of glycogenes to predict the expression of non-glycogenes in each cluster. The normalised weight is represented as the mean weight of all the non-glycogenes predicted by a specific glycogene, multiplied by the number of non-glycogenes that the glycogene could predict. The values were square root transformed to ease

visualisation. **(B)** Functional enrichment of the non-glycogenes predicted in each GEC. **(C)** Drug candidates that inhibit the functional signatures within each GEC.

Tables

Table 1: GHSORM parameters that were used in this study.

	TCGA Luminal A samples (n = 406)	METABRIC Luminal A samples (n = 136)	TCGA (n = 526)	METABRIC (n = 189)
Number of glycogenes used in input layer (SD > 0.5)	277	395	353	277
Representation layer size	250	250	250	250
Number of epochs for denoising autoencoder	1000000	1000000	1000000	100000
SOM initial learning rate	0.7	0.7	0.7	0.8
SOM final learning rate	0.01	0.01	0.01	0.01
τ_1	0.1	0.1	0.4	0.5
τ_2	0.1	0.1	0.05	0.01

Supplementary Table S1: Composition of GECs by PAM50 and mRNAsi.

Supplementary Table S2: RFES performance compared to other commonly used feature discovery wrappers.

Supplementary Table S3: Correlation between glycopaths and immune cell type abundance. Only entries with Pearson coefficient > 0.4 are shown.

Supplementary Table S4: Variable importance of the 60 glycogenes in the best-performing RFES model.

References

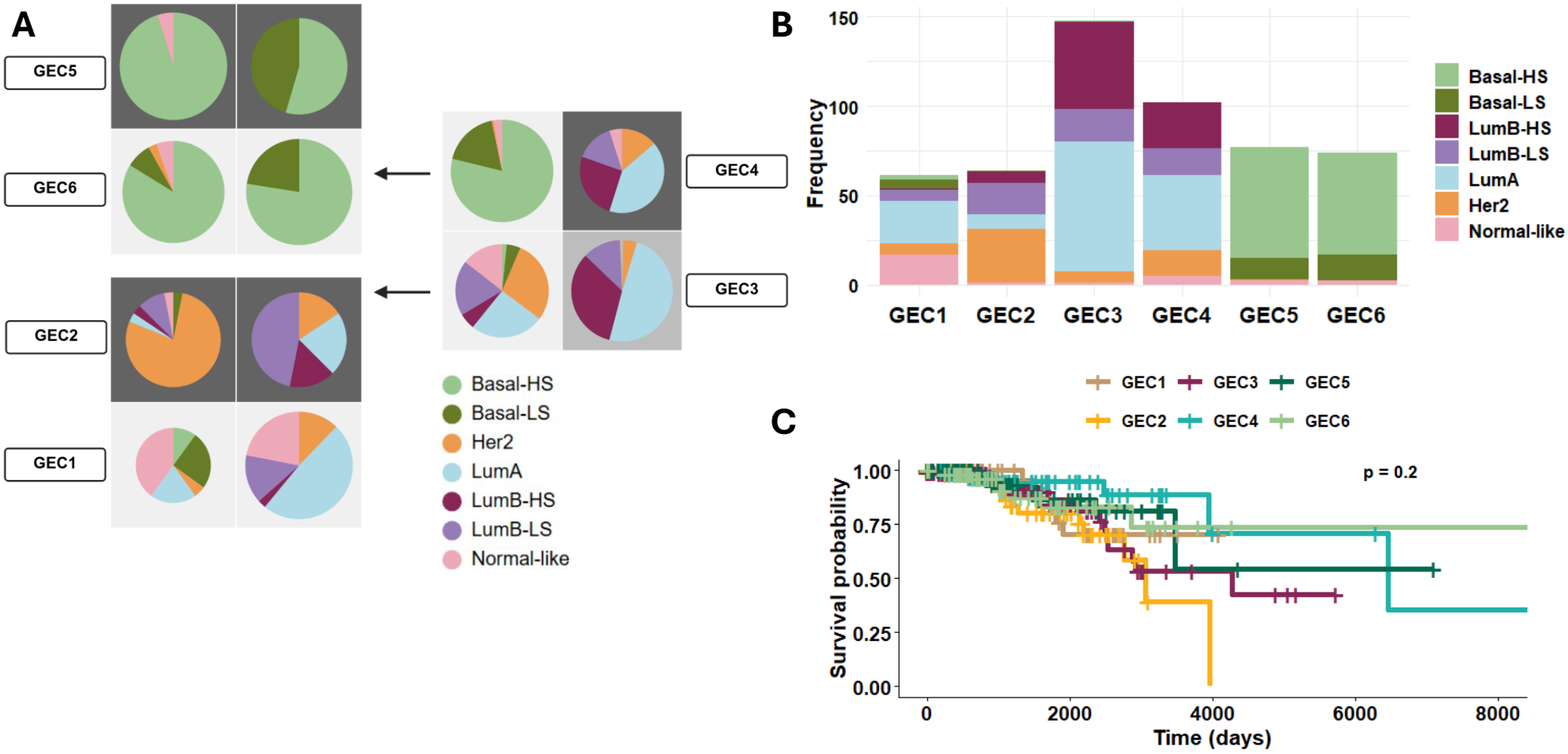
1. Mramor M, Leban G, Demšar J, Zupan B, editors. Conquering the Curse of Dimensionality in Gene Expression Cancer Diagnosis: Tough Problem, Simple Models. Artificial Intelligence in Medicine; 2005 2005//; Berlin, Heidelberg: Springer Berlin Heidelberg.
2. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Medicine. 2021;13:1–17.
3. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC medical genomics. 2015;8:1–14.
4. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology. 2009;27(8):1160–7.
5. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell. 2018;173(2):338–54. e15.
6. Szymiczek A, Lone A, Akbari MR. Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. Clinical genetics. 2021;99(5):613–37.
7. Pinho SS, Reis CA. Glycosylation in cancer: mechanisms and clinical implications. Nat Rev Cancer. 2015;15(9):540–55.
8. Reily C, Stewart TJ, Renfrow MB, Novak J. Glycosylation in health and disease. Nature Reviews Nephrology. 2019;15(6):346–66.
9. Zhang J, Ten Dijke P, Wuhrer M, Zhang T. Role of glycosylation in TGF-β signaling and epithelial-to-mesenchymal transition in cancer. Protein & cell. 2021;12(2):89–106.
10. Boelaars K, van Kooyk Y. Targeting myeloid cells for cancer immunotherapy: Siglec-7/9/10/15 and their ligands. Trends in Cancer. 2024;10(3):230–41.
11. Paszek MJ, DuFort CC, Rossier O, Bainer R, Mouw JK, Godula K, et al. The cancer glycocalyx mechanically primes integrin-mediated growth and survival. Nature. 2014;511(7509):319–25.
12. Thomas D, Rathinavel AK, Radhakrishnan P. Altered glycosylation in cancer: A promising target for biomarkers and therapeutics. Biochim Biophys Acta Rev Cancer. 2021;1875(1):188464.

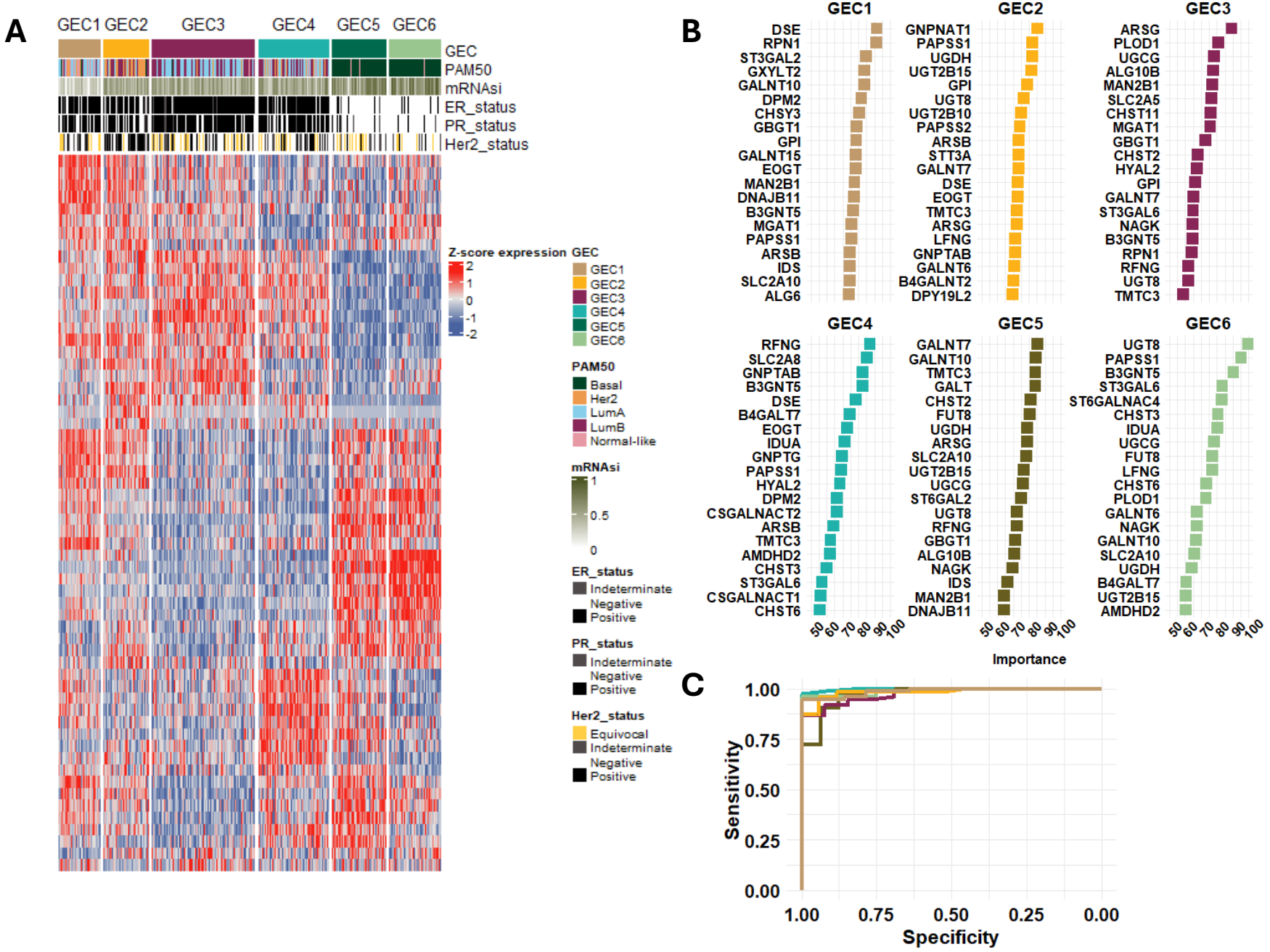
13. Wang Y-C, Lin V, Loring JF, Peterson SE. The 'sweet' spot of cellular pluripotency: protein glycosylation in human pluripotent stem cells and its applications in regenerative medicine. Expert Opinion on Biological Therapy. 2015;15(5):679–87.
14. Nashed A, Dilsook K, Senapathi T, Naidoo K. An in vitro one-pot synthetic biology approach to simulating diverging Golgi O-glycosylation of tumor-associated MUC1 from normal tissue MUC1. 2024.
15. Rabouille C, Hui N, Hunte F, Kieckbusch R, Berger EG, Warren G, et al. Mapping the distribution of Golgi enzymes involved in the construction of complex oligosaccharides. J Cell Sci. 1995;108 (Pt 4):1617–27.
16. Gill DJ, Chia J, Senewiratne J, Bard F. Regulation of O-glycosylation through Golgi-to-ER relocation of initiation enzymes. The Journal of cell biology. 2010;189(5):843–58.
17. Ashkani J, Naidoo KJ. Glycosyltransferase Gene Expression Profiles Classify Cancer Types and Propose Prognostic Subtypes. Scientific Reports. 2016;6(1):26451.
18. Bashir D, Montañez GD, Sehra S, Segura PS, Lauw J, editors. An information-theoretic perspective on overfitting and underfitting. Australasian Joint Conference on Artificial Intelligence; 2020: Springer.
19. Coulson M, Ferles C, Winberg S, Naidoo KJ. Growing Hierarchical Self-Organising Representation Map (GHSORM). Information Sciences. 2023;642:119121.
20. Karim MR, Islam T, Shajalal M, Beyan O, Lange C, Cochez M, et al. Explainable ai for bioinformatics: Methods, tools and applications. Briefings in bioinformatics. 2023;24(5):bbad236.
21. Qi Y. Random forest for bioinformatics. Ensemble machine learning: Methods and applications. 2012:307–23.
22. Strehl A, Ghosh J. Cluster ensembles---a knowledge reuse framework for combining multiple partitions. Journal of machine learning research. 2002;3(Dec):583–617.
23. Ferles C, Papanikolaou Y, Naidoo KJ. Denoising autoencoder self-organizing map (DASOM). Neural Networks. 2018;105:112–31.
24. Chen H, King FJ, Zhou B, Wang Y, Canedy CJ, Hayashi J, et al. Drug target prediction through deep learning functional representation of gene signatures. Nature communications. 2024;15(1):1853.
25. Anderson WF, Rosenberg PS, Prat A, Perou CM, Sherman ME. How many etiological subtypes of breast cancer: two, three, four, or more? Journal of the National Cancer Institute. 2014;106(8):dju165.
26. Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. Computers in Biology and Medicine. 2022;149:106043.
27. Marth JD, Grewal PK. Mammalian glycosylation in immunity. Nature Reviews Immunology. 2008;8(11):874–87.
28. Sica A, Bronte V. Altered macrophage differentiation and immune dysfunction in tumor development. The Journal of clinical investigation. 2007;117(5):1155–66.
29. Liu J, Geng X, Hou J, Wu G. New insights into M1/M2 macrophages: key modulators in cancer progression. Cancer cell international. 2021;21(1):389.

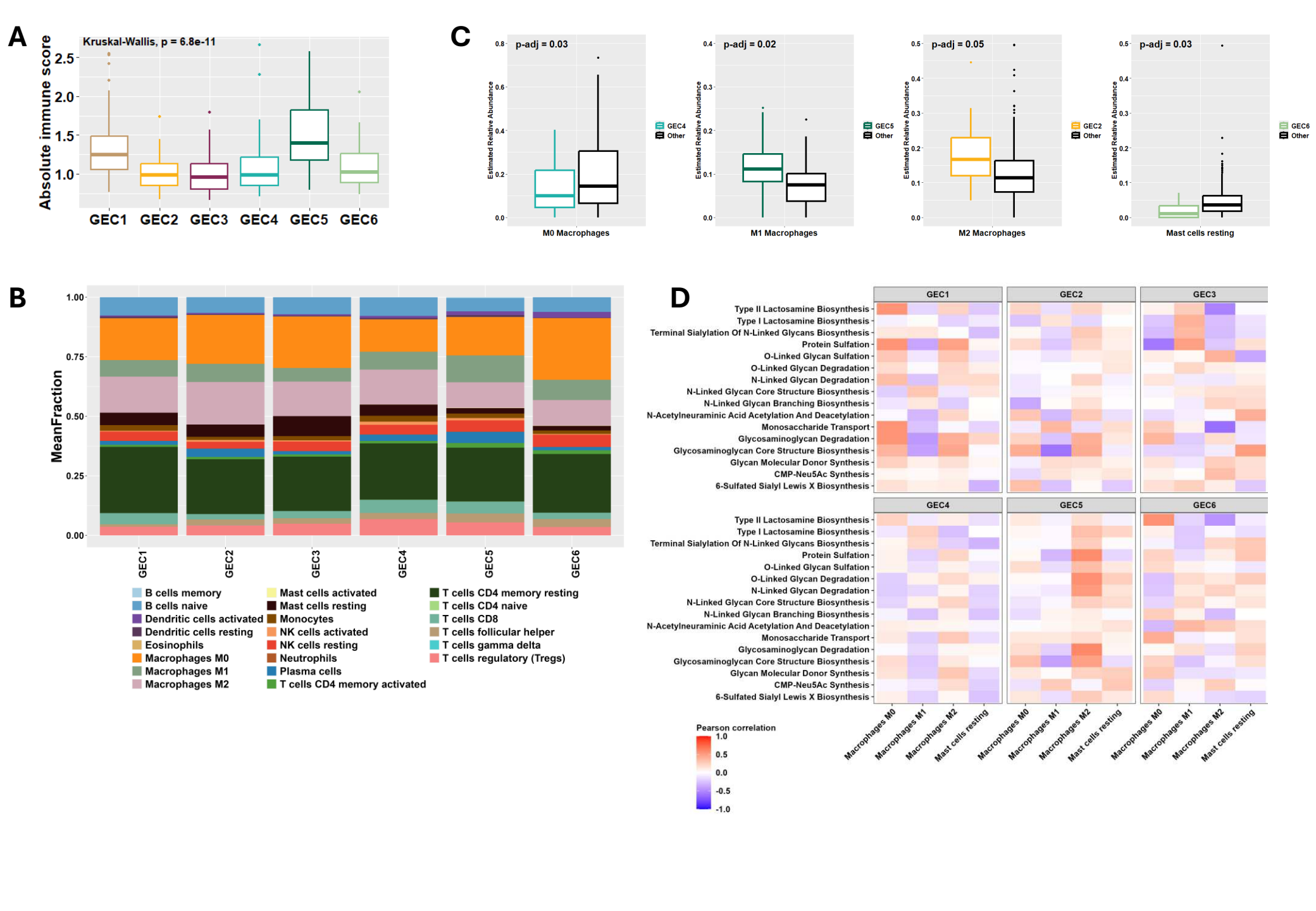
30. Bochner BS. Siglec-8 on human eosinophils and mast cells, and Siglec-F on murine eosinophils, are functionally related inhibitory receptors. Clinical & Experimental Allergy. 2009;39(3):317–24.
31. Zagami P, Carey LA. Triple negative breast cancer: Pitfalls and progress. NPJ breast cancer. 2022;8(1):95.
32. Kurcon T, Liu Z, Paradkar AV, Vaiana CA, Koppolu S, Agrawal P, et al. miRNA proxy approach reveals hidden functions of glycosylation. Proceedings of the National Academy of Sciences. 2015;112(23):7327–32.
33. Reijmers RM, Troeberg L, Lord MS, Petrey AC. Proteoglycans and glycosaminoglycan modification in immune regulation and inflammation. Frontiers Media SA; 2020. p. 595867.
34. Rahman N, Zafar H, Sheikh S, Jabeen A, Choudhary MI. Drug repurposing for the identification of new Bcl-2 inhibitors: In vitro, STD-NMR, molecular docking, and dynamic simulation studies. Life Sciences. 2023;334:122181.
35. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic acids research. 2016;44(8):e71–e.
36. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, Vienna, Austria; 2022.
37. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. 2022.
38. Groth T, Diehl AD, Gunawan R, Neelamegham S. GlycoEnzOnto: a GlycoEnzyme pathway and molecular function ontology. Bioinformatics. 2022;38(24):5413–20.
39. Wickham H, Averick M, Bryan J, Chang W, McGowan LDA, Francois R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019;4:1686.
40. Hocking T. data.table. 2024.
41. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics. 2019;35(12):2084–92.
42. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12).
43. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.
44. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery. 2012;2(5):401–4.
45. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling. 2013;6(269):pl1–pl.
46. Gendoo DM, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2016;32(7):1097–9.
47. Julien Boelaert EO, Jan Sodoge, Mohamed Megdoud, Otmane Naji, Arnaud Lemba Kote, Theo Renoud, Samuel Hym. aweSOM: Interactive Self-Organizing Maps.
48. Wehrens R, Buydens LMC. Self- and Super-organizing Maps in R: The kohonen Package. Journal of Statistical Software. 2007;21(5):1 – 19.

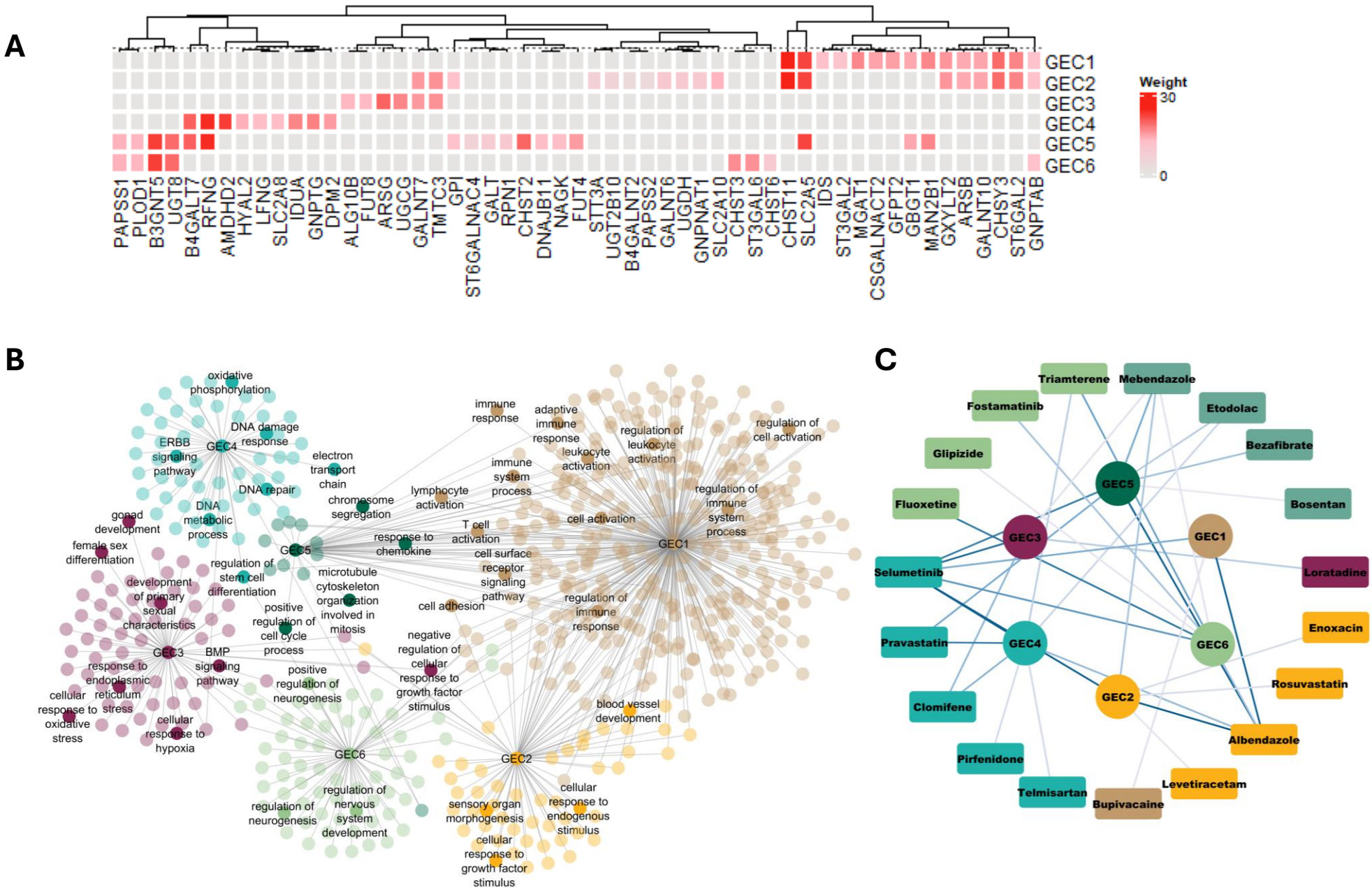
49. Ishwaran H, Lu M, Kogalur UB. randomForestSRC : randomForestSRC algorithm vignette. 2021 [Available from: https://www.randomforestsrc.org/articles/algorithm.html.
50. Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.
51. Kuhn M. The caret Package 2019 [Available from: https://topepo.github.io/caret/index.html#.
52. Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008;28:1–26.
53. Izrailev S. tictoc. 2024.
54. Das B, Krishnan NC, Cook DJ. RACOG and wRACOG: Two probabilistic oversampling techniques. IEEE transactions on knowledge and data engineering. 2014;27(1):222–34.
55. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature biotechnology. 2019;37(7):773–82.
56. van den Boogaart K, Tolosana-Delgado R. compositions: Compositional Data Analysis. 2025.
57. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the national academy of sciences. 2005;102(43):15545–50.
58. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC bioinformatics. 2013;14(1):7.
59. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019;35(12):2159–61.
60. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The innovation. 2021;2(3).
61. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology. 2012;16(5):284–7.
62. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25(8):1091–3.
63. Pantziarka P, Verbaanderd C, Sukhatme V, Capistrano IR, Crispino S, Gyawali B, et al. ReDO_DB: the repurposing drugs in oncology database. ecancermedicalscience. 2018;12.
64. Therneu T. A Package for Survival Analysis in R. 2023.
65. Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, et al. GGally: Extension to 'ggplot2'. 2021.
66. Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using 'ggplot2'. 2021.
67. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847–9.
68. Sakai R, Winand R, Verbeiren T, Moere AV, Aerts J. dendsort: modular leaf ordering methods for dendrogram representations in R. F1000Research. 2014;3.
69. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30(19):2811–2.
70. A. K. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. 2023.

71. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12:1–8.
72. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature genetics. 2000;25(1):25–9.
73. Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023;224(1):iyad031.
74. Fabregat A, Sidiropoulos K, Viteri G, Marin-Garcia P, Ping P, Stein L, et al. Reactome diagram viewer: data structures and strategies to boost performance. Bioinformatics. 2018;34(7):1208–14.
75. Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al. The reactome pathway knowledgebase 2024. Nucleic acids research. 2024;52(D1):D672–D8.
76. Safran M, Rosen N, Twik M, BarShir R, Stein TI, Dahary D, et al. The genecards suite. Practical guide to life science databases. 2021:27–56.
77. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Current protocols in bioinformatics. 2016;54(1):1.30. 1–1.. 3.


A
GEC5
GEC6
GEC2
GEC1
GEC4
GEC3
Basal-HS
Basal-LS
Her2
LumA
LumB-HS
LumB-LS
Normal-like
B
Frequency
150
100
50
0
GEC1
GEC2
GEC3
GEC4
GEC5
GEC6
Basal-HS
Basal-LS
LumB-HS
LumB-LS
LumA
Her2
Normal-like
C
GEC1
GEC3
GEC5
GEC2
GEC4
GEC6
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.2
0
2000
4000
6000
8000
Time (days)


A
GEC1 GEC2 GEC3 GEC4 GEC5 GEC6
GEC
PAM50
mRNAsi
ER_status
PR_status
Her2_status
Z-score expression
2
1
0
-1
-2
GEC
GEC1
GEC2
GEC3
GEC4
GEC5
GEC6
PAM50
Basal
Her2
LumA
LumB
Normal-like
mRNAsi
1
0.5
0
ER_status
Indeterminate
Negative
Positive
PR_status
Indeterminate
Negative
Positive
Her2_status
Equivocal
Indeterminate
Negative
Positive
B
GEC1
DSE
RPN1
ST3GAL2
GXYLT2
GALNT10
DPM2
CHSY3
GBGT1
GPI
GALNT15
EOGT
MAN2B1
DNAJB11
B3GNT5
MGAT1
PAPSS1
ARSB
IDS
SLC2A10
ALG6
GEC2
GNPNAT1
PAPSS1
UGDH
UGT2B15
GPI
UGT8
UGT2B10
PAPSS2
ARSB
STT3A
GALNT7
DSE
EOGT
TMTC3
ARSG
LFNG
GNPTAB
GALNT6
B4GALNT2
DPY19L2
GEC3
ARSG
PLOD1
UGCG
ALG10B
MAN2B1
SLC2A5
CHST11
MGAT1
GBGT1
CHST2
HYAL2
GPI
GALNT7
ST3GAL6
NAGK
B3GNT5
RPN1
RFNG
UGT8
TMTC3
GEC4
RFNG
SLC2A8
GNPTAB
B3GNT5
DSE
B4GALT7
EOGT
IDUA
GNPTG
PAPSS1
HYAL2
DPM2
CSGALNACT2
ARSB
TMTC3
AMDHD2
CHST3
ST3GAL6
CSGALNACT1
CHST6
GEC5
GALNT7
GALNT10
TMTC3
GALT
CHST2
FUT8
UGDH
ARSG
SLC2A10
UGT2B15
UGCG
ST6GAL2
UGT8
RFNG
GBGT1
ALG10B
NAGK
IDS
MAN2B1
DNAJB11
GEC6
UGT8
PAPSS1
B3GNT5
ST3GAL6
ST6GALNAC4
CHST3
IDUA
UGCG
FUT8
LFNG
CHST6
PLOD1
GALNT6
NAGK
GALNT10
SLC2A10
UGDH
B4GALT7
UGT2B15
AMDHD2
50 60 70 80 90 100
Importance
C
Sensitivity
1.00
0.75
0.50
0.25
0.00
1.00 0.75 0.50 0.25 0.00
Specificity


A
Kruskal-Wallis, p = 6.8e-11
Absolute immune score
GEC1
GEC2
GEC3
GEC4
GEC5
GEC6
C
p-adj = 0.03
M0 Macrophages
p-adj = 0.02
M1 Macrophages
p-adj = 0.05
M2 Macrophages
p-adj = 0.03
Mast cells resting
Estimated Relative Abundance
GEC4
GEC5
GEC2
GEC6
Other
B
MeanFraction
B cells memory
B cells naive
Dendritic cells activated
Dendritic cells resting
Eosinophils
Macrophages M0
Macrophages M1
Macrophages M2
Mast cells activated
Mast cells resting
Monocytes
NK cells activated
NK cells resting
Neutrophils
Plasma cells
T cells CD4 memory activated
T cells CD4 memory resting
T cells CD4 naive
T cells CD8
T cells follicular helper
T cells gamma delta
T cells regulatory (Tregs)
D
Type II Lactosamine Biosynthesis
Type I Lactosamine Biosynthesis
Terminal Sialylation Of N-Linked Glycans Biosynthesis
Protein Sulfation
O-Linked Glycan Sulfation
O-Linked Glycan Degradation
N-Linked Glycan Degradation
N-Linked Glycan Core Structure Biosynthesis
N-Linked Glycan Branching Biosynthesis
N-Acetylneuraminic Acid Acetylation And Deacetylation
Monosaccharide Transport
Glycosaminoglycan Degradation
Glycosaminoglycan Core Structure Biosynthesis
Glycan Molecular Donor Synthesis
CMP-Neu5Ac Synthesis
6-Sulfated Sialyl Lewis X Biosynthesis
Macrophages M0
Macrophages M1
Macrophages M2
Mast cells resting
Pearson correlation


A
GEC1
GEC2
GEC3
GEC4
GEC5
GEC6
Weight
30
0
PAPSS1
PLOD1
B3GNT5
UGT8
B4GALT7
RFNG
AMDHD2
HYAL2
LFNG
SLC2A8
IDUA
GNPTG
DPM2
ALG10B
FUT8
ARSG
UGCG
GALNT7
TMTC3
GPI
ST6GALNAC4
GALT
RPN1
CHST2
DNAJB11
NAGK
FUT4
ST3A
UGT2B10
B4GALNT2
PAPSS2
GALNT6
UGDH
GNPNAT1
SLC2A10
CHST3
ST3GAL6
CHST6
CHST11
SLC2A5
IDS
ST3GAL2
MGAT1
CSGALNACT2
GFPT2
GBGT1
MAN2B1
GXYLT2
ARSB
GALNT10
CHSY3
ST6GAL2
GNPTAB
B
oxidative phosphorylation
DNA damage response
ERBB signaling pathway
GEC4
electron transport chain
DNA repair
chromosome segregation
DNA metabolic process
gonad development
female sex differentiation
GEC5
regulation of stem cell differentiation
development of primary sexual characteristics
microtubule cytoskeleton organization involved in mitosis
positive regulation of cell cycle process
response to chemokine
GEC3
BMP signaling pathway
response to endoplasmic reticulum stress
cellular response to oxidative stress
cellular response to hypoxia
positive regulation of neurogenesis
negative regulation of cellular response to growth factor stimulus
GEC6
regulation of neurogenesis
regulation of nervous system development
immune response
adaptive immune response
regulation of leukocyte activation
regulation of cell activation
leukocyte activation
immune system process
lymphocyte activation
T cell activation
cell activation
regulation of immune system process
GEC1
cell surface receptor signaling pathway
cell adhesion
regulation of immune response
blood vessel development
GEC2
sensory organ morphogenesis
cellular response to endogenous stimulus
cellular response to growth factor stimulus
C
Triamterene
Mebendazole
Fostamatinib
Etodolac
Glipizide
Bezafibrate
GEC5
Fluoxetine
Bosentan
GEC3
GEC1
Selumetinib
Loratadine
Pravastatin
GEC4
GEC6
Enoxacin
Rosuvastatin
GEC2
Clomifene
Albendazole
Pirfenidone
Levetiracetam
Telmisartan
Bupivacaine

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.xlsx
- SupplementaryTableS2.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS4.xlsx
- SupportingInformationAdvDigChemDiscF.pdf