A systems based framework to deduce transcription factors and signaling pathways regulating glycan biosynthesis

Theodore Groth[1] and Sriram Neelamegham[1,2,3,*]

[1]Chemical and Biological Engineering, [2]Biomedical Engineering and [3]Medicine
University at Buffalo, State University of New York, Buffalo, NY 14260, USA

Running title: Systems Glycobiology

*** Correspondence:** Sriram Neelamegham, 906 Furnas Hall, Buffalo, NY, 14260, neel@buffalo.edu, Ph: 716-645-1200; Fax: 716-645-3822

Abstract

Glycosylation is a common, complex, non-linear post-translational modification. The biosynthesis of these structures is regulated by a set of 'glycogenes'. The role of transcription factors (TFs) in regulating the glycogenes and related glycosylation pathways is yet unknown. This manuscript presents a multi-OMICs data-mining framework to computationally predict tissue specific TF activities and cell signaling pathways regulating the biosynthesis of specific glycan structures. It combines existing ChIP-Seq (Chromatin ImmunoPrecipitation Sequencing) and RNA-Seq data to reveal 20,617 potentially significant TF-glycogene relationships. This includes interactions involving 524 unique TFs and 341 glycogenes that span 29 TCGA (The Cancer Genome Atlas) cancer types. Here, TF-glycogene interactions appeared in clusters or 'communities', suggesting that they may collectively drive changes in sets of carbohydrate structures rather than unique glycans as disease progresses. Upon applying the Fisher's exact test along with glycogene pathway ontology, we identify TFs that may specifically regulate the biosynthesis of individual glycan types. Integration with knowledge from the Reactome database established the link between cell signaling pathways, transcription factors, glycogene expression, and glycosylation pathways. Whereas analysis results are presented for all 29 cancer types, specific focus is placed on human luminal and basal breast cancer disease progression. This implicates a key role for TGF-β and Wnt signaling in regulating TFs that control both tumorigenesis and cellular glycosylation. Overall, the computational predictions in this manuscript present a rich dataset that is ripe for experimental testing and hypotheses validation.

Keywords: Glycoinformatics, transcription factor, glycosylation, ChIP-Seq, TCGA

Introduction

The glycan signatures of cells and tissue is controlled by the expression pattern of 200-300 glycosylating enzymes that are together termed 'GlycoEnzymes' [1]. The expression of these glycoEnzymes is in turn driven, in part, by the action of a class of proteins called transcription factors (TFs). These TFs regulate gene expression by binding proximal to the promoter regions of genes, facilitating the binding of RNA polymerases. They may homotropically or heterotropically associate with additional TFs in order to directly or indirectly control messenger RNA (mRNA) expression. Among the TFs, some 'pioneer factors' can pervasively regulate gene regulatory circuits, and access chromatin despite it being in a condensed state[2] . These TFs act as 'master regulators', promoting the expression of several genes across many signaling pathways, such as differentiation, apoptosis, and cell proliferation. The precise targets of the TFs is controlled by their tissue-specific expression, DNA binding domains and nucleosome interaction sequences[2] . Additional factors regulating transcriptional activity include: i. cofactors, small molecules or proteins, that enable TF binding to their DNA recognition sites and optimal RNA polymerase recruitment[2] ; ii. chomatin modifications, such as acetylation, methylation and phosphorylation, which alter TF access to DNA binding segments; and iii. methylation of CpG islands in promoter regions which can inhibit the expression of specific genes[3,4] . To date, the interactions between glycoEnzymes and TFs has not been systematically elucidated[5–7] .

A number of high-throughput experimental methods that use either cell systems or degenerate oligonucleotide libraries can aid the mapping of TFs to glycogene expression. Most common among them is the Chromatin ImmunoPrecipitation Sequencing (ChIP-Seq) technique, where TFs are crosslinked to bound genomic DNA in cells, pulled down using specific antibodies, and then the associated DNA are released and identified using next generation sequencing (NGS) technology[8,9] . In addition to identifying the position of TF binding to the genome, the ChIP-seq data also reveal transcription factor sequence binding specificity. This

binding specificity can be summarized in a position weight matrix (PWM), which captures the likelihood of observing nucleotides at various positions along a DNA sequence. By extension, methylation sites proximal to TF binding sites can be mapped using the bisulfite ChIP-seq method[8] . Since only one TF can be screened in the classical ChIP-Seq workflow, a variation called Re-ChIP has emerged that uses more than one anti-TF antibody to enable the identification of complexes containing multiple TFs[10] . The sequences obtained in a ChIP-seq experiment may be biased depending on the epigenetic state of the cell, as not all binding sites may have been available in the native cell. To overcome this limitation, a set of reductionist approaches have been developed under the umbrella of the Systematic Evolution of Ligands through eXponential Enrichment (SELEX) assay[11] . Here, an unbiased evaluation of TF binding specificity is performed by quantifying the binding of randomized nucleotides from a pool to the TFs. In improvements to this method, multiple TFs complexed with DNA can also be detected using Consecutive Affinity-Purification SELEX (CAP-SELEX), which detects interacting pairs of transcription factors bound to oligonucleotides through tandem-affinity purification[12] . SELEX data, generated in this manner, can then be used to infer TF binding sites throughout a genome. Many datasets generated using the above techniques are now publicly available at the Gene Expression Omnibus (GEO).

In the current manuscript, we sought to utilize a multi-OMICs framework to relate cell-specific signaling processes, transcription factors, glycogenes and glycosylation pathways (**Fig. 1A**). This framework integrated ChIP-Seq and RNA-Seq experimental data with glycosylation pathway ontology and cell signaling knowledge. Here, ChIP-Seq determines a list of target genes bound by specific TFs, including data on proximity to the transcription state site (TSS). However, whether this interaction actually regulates gene expression cannot be inferred based on binding data alone. To address this limitation, data collated at the Cistrome Cancer database were used to determine if there exists a correlation between TF and gene expression. This

database uses TF-gene binding data from previously published ChIP-Seq studies for various cancer cell lines and cancer tissue RNA-seq data from The Cancer Genome Atlas (TCGA)[13] . Thus, the approach establishes a tissue-specific TF-gene expression relationship for 29 RNA-Seq-based cancer types from the TCGA. A subset of these data establish the TF-glycogene relationship. Further analysis of these data using a glycosylation pathway framework available at GlycoEnzDB (unpublished data), yielded predictions of potential TFs contributing to cellular glycosylation pathways and tissue specific glycan signatures. Finally, using the Reactome Database's Overrepresentation API[14] , we established the link between signaling pathways and TFs, thus closing the loop among the multi-OMICs data (**Fig. 1B**). Overall, we propose that this computational framework that links multiple OMICs methods can be used for hypothesis generation and experimental validation.

Results

TF-Glycogene interaction map and relation to cell signaling pathways: The manuscript follows a workflow shown in **Figure 2**. It infers TF-glycogene relationships using publicly available ChIP-seq data and RNA-Seq results from The Cancer Genome Atlas (TCGA). These data were obtained from the curated Cistrome Cancer DB[15] for 524 TF targets. The strength of the relationship of these TFs to 341 glycogenes (**Supplemental Table S1**) was inferred using two metrics: the regulatory potential (RP) which is a measure of TF binding proximity to the gene transcriptional start site; and the Spearman's correlation (ρ) which describes the correlation between TF and target-gene expression. Such analysis was performed for 29 cancer types listed in **Supplemental Table S2**. The analysis revealed 20,617 high-strength TF-glycogene interactions. These can be visualized in the Cytoscape session files for each of the cancers individually (**Supplemental File S1**). Attempts were made to link the TFs identified in these analysis to cell signaling pathways using the Reactome DB overrepresentation API, and glycogenes to specific pathways using knowledge available from GlycoEnzDB. This cancer-specific TF-glycogene interaction analysis revealed communities of co-regulated TFs and glycogenes that may be indicative of concerted biological processes. Using these TF-glycogene data, Robust Rank Aggregation (RRA) metrics were also generated in order to determine TF-glycogene interactions that are commonly regulated among the different cancers. These represent potentially significant molecular interactions that could be tested experimentally.

Next, the Fisher's exact test was used to infer TF-glycogene interactions that may regulate glycosylation pathways. To achieve this, 212 of the glycogenes were classified into 20 glycosylation pathways/groups based on curation at GlycoEnzDB (**Supplemental Table S3**). TF-pathway relationships identified in this manner were related to knowledge available at ReactomeDB. This resulted in a relationship between cell-signaling, TF activity regulation and glycan structure changes (**Supplemental Table S4, S5**). These data are presented as Alluvial plots for the 29 cancer types (**Supplemental Fig. S2**). Here, the TFs were linked to

glycosylation pathways by colored bands if they were found to regulate a disproportionately high fraction of glycogenes belonging to that pathway. Likewise, biological pathways were linked with TFs if that TF was found to be enriched in the biological pathway. Reading these alluvial plots left to right, one can deduce which biological pathways may be involved in regulating TFs, and how this TF could be regulating glycosylation. While detailed TF-glycogene and TF-glycosylation pathway analysis is possible for each of the cancers, this manuscript focused on the TFs that are enriched for luminal and basal forms of breast cancer (discussed below).

TF-glycogene communities in breast cancer (cytoscape plots): Breast cancers appear in 5 unique molecular subtypes based on the PAM50 classification[16] . These include: i. normal-like, ii-iii. luminal A and luminal B which overexpress estrogen receptor ESR1, iv. Her2+ tumors that overexpress the epidermal growth factor receptor (ERBB), and v. basal (triple negative) that express neither ESR1 nor ERBB. Each of these subtypes has unique signaling mechanisms that may contribute to different glycan signatures. Using Reactome DB knowledge, we establish this link between cell signaling, TFs and glycan structures (**Fig. 3**). A detailed discussion based on current knowledge in literature follows.

Luminal breast cancers had three large communities of TF-glycogene interactions based on cytoscape “clusterMaker” analysis[17] . For each community, Reactome DB overrepresentation analysis was performed on the TFs. The largest community detected had TFs enriched for RUNX3 signaling, IL-21 signaling, MECP2, and PTEN regulation (**Fig. 4a**). Overrepresented glycosylation pathways in this community included pathways regulating sialylation, hyaluronan synthesis, and chondroitin and dermatan sulfate elongation. STAT1, 4, and 5 proteins were found to be enriched in the IL-21 signaling pathway. Luminal breast cancers are known to express STAT1, 3 and STATs 2 and 4 are known to be expressed in luminal breast cancer cell lines. STAT5 is known to be constitutively active in luminal breast cancer and confers anti-apoptotic characteristics to cells[18] . The other two communities

detected consisted primarily of chromatin-modifying enzymes. Complex N-linked glycan synthesis and the dolichol pathway were significantly enriched in the second community. In the third community, O-linked mannose and LacdiNAc synthesis were disproportionately regulated. Overall, the pathway maps suggest that chromatin remodeling enzymes could potentially play roles in regulating glycan synthesis in luminal breast cancer. Based on the appearance of communities, groups of glycans would be expected to be simultaneously dysregulated during cancer, and together these may serve as robust indicators of disease progression.

Like luminal, basal breast cancer TF-glycogene relationships were also clustered into three communities. Here, the first community was enriched for chromatin modifying enzymes, with complex N-linked glycan synthesis bring the primary glycosylation pathway being affected (**Fig. 5a**). The second community was enriched for interferon α/β/γ signaling pathways, with interferon regulatory factor (IRF) transcription factors being enriched. The TFs IRF-1 and IRF-5 have been shown to act as tumor suppressors in breast cancer[19,20] . Their loss-of-function event in breast cancer could potentially downregulate O-linked fucosylation. The third community of basal breast cancer did not exhibit any specific TF pathway enrichments.

Linking cell signaling to TF and glycogenes for luminal breast cancer (alluvial plot): The Fisher's exact test was performed to identify TF-glycogene relationships that are enriched in individual glycosylation pathways. This analysis was performed individually for all 29 cancer types. These findings were related to pathway knowledge in the Reactome DB, in order to generate a number of experimentally testable hypotheses. These links between biological signaling pathways, TFs, and glycosylation pathways are shown in alluvial plots for luminal and basal breast cancers (**Fig. 4b, 5b**) , with additional plots provided for additional cancer types in Supplemental Material. Below we discuss our findings for luminal breast cancer (**Fig. 4b**):

1. CREB3L4 and PRDM1 disproportionately affects LacNAc pathway in luminal breast cancer:

We observed that Type 1 and 2 LacNAc were regulated by the transcription factor CREB3L4, which is associated with the CREB3 pathway. CREB molecules act as receptors to cellular stress in hypoxic environments or during protein folding stress. Once the stress is detected in the endoplasmic reticulum (ER) or Golgi apparatus, their cytoplasmic domains cleave and are transported to the nucleus to act as transcription factors[21] . The CREB3L4 molecule is associated with detection of protein folding stress in the ER[22] . This molecule has been found to be upregulated in breast cancer respect to normal. The depletion of this gene results in apoptosis in breast cancer cell lines, suggesting a proliferative effect of CREB3L4[23] . The increase of CREB3L4 expression upon breast cancer development could potentially increase the prevalence of poly LacNAc structures on N- and O-linked glycans, which have been shown to play roles in metastasis[24] . The enzyme responsible for enriching CREB3L4 for these pathways is the B4GALT3 glycogene, which adds galactose in a β1-4 linkage. It is possible that CREB3L4 increases the prevalence of these structures via B4GALT3 based on their statistical metrics: ρ=0.56, RP=0.94. Here, the high RP indicates proximal location between the TF binding site and the transcription start site. The high Spearman correlation value indicates a positive correlation between CREB3L4 and B4GALT3 expression.

Our analysis suggests that, in addition to CREB3L4, the TF PRDM1 may also regulate Type 1and 2 LacNAc extension. This TF, which is also known as Blimp-1, is a transcriptional repressor. PRDM1 is upregulated in breast cancer, and can induce the expression of Snail via intermediate downregulation of other signaling proteins[25] . The glycogene found in this process to regulate Type 1and 2 LacNAc type structures is B3GNT5 (ρ=0.60, R.P.=0.84).

2. E2F1 and MYBL2 disproportionately affect Dolichol synthesis pathway: Our analysis reveals that E2F1 may be a key enzyme regulating the dolichol biosynthesis pathway. This TF is known to be involved in metabolic homeostasis, regulation of cell cycle, and it is activated in response

to DNA damage. Depending on the cofactors associated with E2F1, it may act as a transcriptional repressor or activator. During cancer development, E2F1 has been shown to promote cancer metabolism dysregulation such as promoting the Warburg effect by simultaneously upregulating glycolysis and downregulating oxidative phosphorylation genes[26] . In breast cancer-specific contexts, it has been shown that E2F1 positively regulates metastasis-related genes and promotes mobility[27] . E2F1 regulates the function of two enzymes in the dolichol pathway, ALG3 (ρ=0.43, RP=1.00) and DPM1 (ρ=0.75, RP=0.40). In this regard, ALG3 is responsible for adding mannose to the N-linked precursor structure, and DPM1 is responsible for transferring mannose to dolichol in the outer ER.

Like E2F1, MYBL2 is another TF involved in regulation of cell cycle. It is activated in the G2/early S phase of cellular replication[28] . In cancers, MYBL2 can become amplified through the chromosomal amplification or through the repression of the dimerization partner, RB-like proteins, E2Fs and MuvB core (DREAM) complex, responsible for repressing MYBL2 in quiescent cells. Increased MYBL2 expression in tumors results in cell proliferation, survival, and EMT[28] . In our analysis, ALG3 (ρ=0.50, RP=0.82) and DPM1 (ρ=0.71, RP=0.42) were both responsible for enriching MYBL2 to the dolichol pathway. In addition to dolichol pathway regulation, MYBL2 may also regulate the function of two glucosyltransferases RPN1 (ρ=0.43, RP=0.80) and RPN2 (ρ=0.42, RP=0.42). They are responsible for adding glucose onto the α1-3 mannose branch on the N-linked glycan precursor.

3. MEF2C disproportionately regulates Glycosaminoglycan synthesis pathways: MEF2C was found to regulate several glycogenes in the chondroitin and dermatan sulfate synthesis pathways. This TF plays roles in development, particularly with the development of neurons and hematopoetic cell differentiation towards myeloid lineages. It has been found that MEF2C can be upregulated in several cancer types such as myeloid leukemia, immature T-cell acute lymphoblastic leukemia, and rhabdomyosarcoma[29] . It is known that MEF2C is directly

impacted by TGF-β signaling, thus increasing metastatic potential of cancer[30] . MEF2C was found to be inhibited by MECP2 based on the Reactome pathway enrichment. Since the glycosaminoglycan elongation pathways positively correlate to MEFC2 expression, and MEFC2 is amplified in cancer, it is possible that MECP2 may not sufficiently expressed to repress MEFC2 in call cancer cells. Glycosyltransferases responsible for enriching MEF2C to the GAG synthesis pathways include CSGALNACT1 (ρ=0.66, RP=0.71), CHST3 (ρ=0.50, RP=0.74), CHST11 (ρ=0.47, RP=0.84), DSEL (ρ=0.40, RP=0.81), and UST (ρ=0.42, RP=0.95). Here, CSGALNACT1 is responsible for the addition of GalNAc to glucuronic acid to increase chondroitin polymer length, CHST3, CHST11, and UST are involved in the sulfation of GalNAc and iduronic acid, and DSEL is the epimerase which converts glucuronic acid to iduronic acid in CS/DS chains.

4. MECP2 and SMAD4 disproportionately regulated heparan sulfate chain elongation: The Methyl CPG binding Protein 2 (MECP2) transcription factor was found to positively regulate heparan sulfate elongation. MECP2 regulates gene expression by binding to methylated promoters, and then by recruiting chromatin remodeling proteins to condense DNA and repress gene expression[31,32] . In breast cancer, it is thought that MEPC2 inhibits the p53 pathway via the epigenetic upregulation of RPL5 and RPL11, thus causing cancer proliferation[33] . Additionally, it participated in promoting ERK1/2 signaling in breast cancer[34] . The glycogene NDST1 (ρ=0.41, RP=0.67) was responsible for enriching MECP2 to the heparan sulfate elongation pathway. This enzyme is a sulfotransferase that sulfates N-acetyl glucuronic acid in heparan polymers.

SMAD4 is a transcription factor directly regulated by TGF-β signaling. SMAD4 must complex with the SMAD2/3 dimer before it acts as a functional transcription factor complex in the nucleus[35] . SMAD4 acts as a tumor suppressor in breast cancer contexts. Downregulation of SMAD4 in the triple negative breast cancer cell line MDA-MB-231 induces TGF-β -driven

EMT, and also facilitates the metastasis of tumors to bone[36] . Typically, SMAD4 expression is much lower in breast tumor tissue compared with adjacent normal tissues[37] . The GLCE glycogene (ρ=0.44, RP=0.63) enriched SMAD4 to heparan sulfate synthesis, and is responsible for converting glucuronic acid to iduronic acid.

5. TCF7L2 disproportionately regulates sialic acid glycosyltransferases: Transcription factor 7-like 2 (TCF7L2) is regulated by Wnt β-catenin signaling. β-catenin complexes with TCF7L2 upon translocation into the nucleus to initiate transcription[38] . This TF is important in gluconeogenesis in the liver, adipogenesis, regulation of hormone synthesis, and pancreas homeostasis[39] . TCF7L2 exhibits polymorphisms which results in loss-of-function, and can promote metastatic phenotypes in colorectal cancer[40] . Polysialylation glycogenes ST8SIA1 (ρ=0.43, RP=0.65) and ST8SIA2 (ρ=0.43, RP=0.95) were enriched to the sialylation pathway were associated with TCF7L2 regulation. Both are involved in the polysialylation of glycosphingolipids.

Linking cell signaling to TF and glycogenes for basal breast cancer (alluvial plot): Fewer transcription factors were found to be enriched to pathways in basal breast cancer compared to luminal cancer (**Fig. 5b**). The roles of the enriched TFs and their relation to glycogenes and cancer is elaborated below.

1. Critical role for RUNX3 in terminal fucosylation: The terminal fucosyltransferase FUT7 (ρ=0.49, RP=0.89) was found to be positively regulated by the RUNX3 TF. The RUNX family of transcription factors (including RUNX1-3), are involved in several developmental processes, including hematopoiesis, immune cell activation, and skeletal development. It was discovered that RUNX3 acts as a tumor suppressor gene in breast cancer, as well as others. Here, hypermethylation of RUNX3 leads to reduction in TF activity and loss of tumor suppression

activity[41] . Our data suggest that this may be associated with a reduction of FUT7 activity thus impacting the expression of the sialyl Lewis-X antigens in basal tumors.

2. Regulation of O-glycosylation by SMAD2: SMAD2 was found to significantly affect core 1 & 2 O-linked glycan structures. SMAD proteins are activated by TGF-β signaling and bind to DNA to act as cofactors to recruit TFs. SMAD2 is one of the receptor-regulated SMADs (R-SMAD), meaning that it is directly phosphorylated by the TGF-β receptor. Once phosphorylated, it must bind to the common partner SMAD (Co-SMAD, SMAD4) to gain entry into the nucleus. The Co-SMAD R-SMAD complex binds DNA and recruits TFs to regulate gene expression. Breast cancers have increased proliferation upon cancer development when the R-SMAD molecules are dysregulated. It has been shown that overexpression of SMAD3 in breast cancer cell lines can increase proliferative signaling in the normal breast cell line MCF10A, however it did not have an effect on EMT markers[42] . Another experiment showed that downregulating SMAD2 in the basal breast cancer cell line MDA-MB-231 increased cell proliferation and metastatic potential to bone[43] . Thus, SMAD2 acts as a tumor metastasis suppressor. This TF was found to regulate GALNT1 (ρ=0.54, RP=1.00), which adds GalNAc to serine or threonine residues to being core 1 and 2 O-linked glycan synthesis. Thus, SMAD2 may play a key role in regulating Tn-antigen expression in proteins like MUC-1 that are associated with breast cancer progression.

Transcription factors broadly affecting glycosylation: Robust Rank Aggregation (RRA) was applied to determine TFs that may broadly regulate glycosylation across all cancer types (**Supplemental Table S6**). Given ranked lists based on *RP* and Spearman's ρ, RRA statistically evaluated whether a feature has a high ranking across all lists. Such analysis was performed for individual glycosylation pathway, independently. The top-10 enriched TFs is shown in **Fig 6**. Some pathways had TFs with much lower RRA statistics that others, including chondroitin and

dermatan sulfate extension, complex N-linked glycan formation, dolichol pathway, glycolipid core synthesis, N-linked glycan branching, and O-linked fucose.

JMJD1C, a histone demethylase[44] , regulated chondroitin and dermatan sulfate extension by its interaction with two GalNAc transferases: CSGALNACT2 and CHSY1. It also interacts with genes regulating GAG initiation. This gene is known to promote colon cancer metastasis through ATF2 interactions[45] . JMJD1C regulates these genes with high regulatory potential and with high correlation in 13 different cancer types (Luminal and Basal BRCA, COAD_READ, GBM, HNSC, KICH, KIRC, LGG, LUAD, MESO, PAAD, PGPC, PRAD, SARC, and THYM). This analysis suggests an candidate epigenetic regulator of glycosaminoglycan biosynthesis across cancer types.

Mediator subunit 1 (MED1) is a transcriptional cofactor known to comprise enhancer complexes, and is regulated by hormone signals in breast, prostate, and bladder cancers[46,47] . For example, it mediates the binding between the estrogen receptor and a mediator complex responsible for recruiting RNA polymerase II in breast cancer complexes[47] . This TF was enriched to complex N-linked glycan synthesis pathways through the regulation of GANAB. This relationship is present in 13 cancer types (BLCA, CESC, COAD_READ, HNSC, KIRP, LGG, LUAD, MESO, PAAD, PRAD, TGCT, THYM, and UCEC).

PRDM4 was found to be enriched to N-linked glycan branching. This protein has been shown to induce EMT via YAP1-mediated transcriptional regulation to upregulate IGTB1[48] . This gene regulates MGAT5, an important glycogene for creating branched N-linked glycans that play roles in metastasis. It may be possible that PRDM4 may be promoting two mechanisms of metastasis simultaneously. MGAT5 is regulated in 12 different cancer types (BLCA, COAD_REA, KIRC, KIRP, LAML, LGG, MESO, PCPG, PRAD, TGCT, THCA, UCEC).

One TF, LYL1, shows the potential to regulate many glycosylation pathways simultaneously across cancer types. This protein has been shown to interact with CREB1, and may be involved in cellular stress maintenance[49] . This TF was in the top 10 most enriched

TFs for chondroitin and dermatan sulfate extension, fucosylation, ganglioside synthesis, sulfated glycan epitopes, sialylation, O-linked fucose. It was found to regulate 57 different glycogenes across 22 cancer types. Further knowledge as to which cofactors associate with LYL1 or CREB1 may provide knowledge as to how LYL1 regulates these genes.

Discussion

In the current analysis, we sought to identify strategies to enhance systems glycobiology knowledge by leveraging existing high-throughput gene expression data, specifically publicly available ChIP-Seq and RNA-Seq datasets. As an example, we present a framework for the identification of TFs regulating glycogenes and glycosylation processes in 29 different cancer types. This analysis reveals 20,617 potentially significant TF-glycogene across the 29 cancer types. Approximately three glycogenes were regulated by a given TF based on our filtering criteria, with this number ranging from 1-10. These findings are tissue-specific, as TF and glycogene expression vary widely among the different cell types. The analysis also revealed putative TF-glycogene interactions that disproportionately impact specific glycosylation pathways. Knowing which TF regulates which glycogene and pathway in a context-dependent manner can provide insight as to how signaling pathways contribute to altered glycan structures in diseases such as diabetes and cancer. Thus, this work represents a rich starting point for wet lab validation and glycoinformatics database construction.

Visualizing TF-glycogene interaction networks revealed communities of glycogenes in each cancer type. The presence of chromatin-modifying enzymes in large regulatory communities in both luminal and basal breast cancer suggests a role of epigenetics in glycogene regulation. To date, a systems-level investigation evaluating the epigenetic states of cell systems on the resulting glycome has not been performed. Our results suggest that complex N-linked branching and glycosylation may be sensitive to these processes. The signaling pathways enriched in the largest community in luminal breast cancer were reflected in our pathway enrichment findings. RUNX3, interleukin signaling, and the involvement of MECP2 regulation were all found to disproportionately regulate sialic acid and GAG synthesis pathways. Several of the TFs enriched to glycosylation pathways were either regulated by or involved in TGF-β signaling and Wnt β-catenin signaling. These TFs primarily affected glycosaminoglycan synthesis pathways, sialylation and Type-2 LacNAc synthesis. Some of these glycan structures

have been implicated in the metastasis of tumors[50,51] . Cell cycle and metabolic regulatory TFs were shown to regulate some glycogenes involved in the dolichol pathway. The crosstalk between cell cycle and glycosylation is not well explored, and could potentially be important for understanding N-linked glycosylation flux in cancer. Some TFs were found to interact with methyl CpG binding TFs when regulating glycosaminoglycan proteins, implicating methylation as a possible modulator of glycosylation in cancer.

The framework described in this manuscript represents a starting point for the development of new glycoinformatics methods, using readily available NGS datasets. Perturbing the values of RP and ρ, coupled with RRA, may reveal prevailing TF-glycogene-pathway relationships that are not sensitive to the selection of data analysis parameters. Additionally multi-OMICs data mining could also aid validation. These include: i. Other ChIP-Seq databases, such as the Gene Transcription Regulatory Database (GTRD)[52] that have analyzed vast publicly available ChIP-Seq data to systematically cataloged TF-gene relationships across several organisms and cellular contexts. ii. the Regulatory Circuits[53] database that quantify the activity of promoters and enhancer regions through cap analysis of gene expression (CAGE) and expression Quantitative Trait Loci (eQTL), respectively. Finally, wet-lab experiments quantifying the effect of TF knockout/overexpression on glycogenes using single-cell RNA-Seq may provide evidence of TF-glycogene relationships. By extension, complementary mass spectrometry based glycomics and glycoproteomics studies could strengthen conclusions regarding cell signaling-TF-pathway relationships.

The current analysis systematically describes putative connections between TF regulation and glycosylation pathway activity in 29 cancer types. It reveals that EMT-driving pathways, such as TGF-β and Wnt β-catenin signaling, can drive concerted changes in several glycan classes. These alterations appear in communities, and may collectively drive clinically detected cancer regulators and glycan disease biomarkers.

EXPERIMENTAL METHODS

Glycogene-pathway classification: A list of 212 unique glycogenes involved in 20 different glycosylation pathways were used in this work (**Supplemental Table S3**). These data are collated from GlycoEnzDB (virtualglycome.org/GlycoEnzDB), with original data coming from various sources in literature[54,55] . The following is a summary of the pathways studied and the enzymes involved:

1. Glycolipid core: The enzymes in this group are involved in the biosynthesis of the glucosyl-ceramide (GlcCer) and galactosyl (GalCer)-ceramide lipid core. Here, the GlcCer core is formed by the UDP-glucose:ceramide glucosyltransferase (UGCG) which transfers the first glucose. Following this, lactosylceramide is formed by the action of the β1,4GalT activity of B4GalT5 (and possibly also B4GalT3, 4 and 6). The GalCer core is typically structurally small and is made by UDP-Gal:ceramide galactosyltransferase (UGT8). These structures can be further sulfated by GAL3ST1 or sialylated by ST3GAL5.

2. P1-Pk Blood Group: The Pk, P1 and P antigens are synthesized on lactosyl-ceramide glycolipid core. The activity of α1-4GalT (A4GALT) on this core results in the Pk antigen, followed by β1-3GalNAcT (B3GALNT1) to form the P antigen. The P1 antigen, on the other hand, is formed by the sequential action of β1-3GlcNAcT (B3GNT5), β1-4GalT (B4GALT1-6) and α1-4GalT (A4GALT) on the glycolipid core.

3. Gangliosides: This pathway encompasses all glycogenes responsible for synthesizing a/b/c gangliosides. UGCG is included to consider the addition of glucose to ceramide. ST3GAL5, and ST8SIA enzymes are added to take the core ganglioside structures to the a,b and c levels. B4GALTs and B4GALNT1 are included to account for ganglioside elongation. Decoration of the gangliosides with sialic acid occurs using ST6GALNAC3-6 and also ST8SIA1/3/5.

4. Dolichol Pathway: This results in the formation of the dolichol-linked 14-monosaccharide precursor oligosaccharide. This glycan is co-translationally transferred en bloc onto Asn-X-Ser/Thr sites of the newly synthesized protein as it enters the endoplasmic reticulum. The

enzymes involved is such synthesis include the ALG (Asparagine-linked N-glycosylation) enzymes, and additional proteins (part of OSTA and OSTB) involved in the transfer of the glycan to the nascent protein.

5. Complex N-glycans: This pathway includes glycogenes responsible for processing the N-linked precursor structure emerging from the dolichol pathway into complex structures. Enzymes involved include mannosidases, glucosidases, some enzymes facilitating protein folding and also enzymes that direct acid hydolases to the lysosome.

6. N-glycan branching: These glycogenes are responsible for the addition of GlcNAc to processed N-linked glycan structures. These include all the MGAT enzymes.

7. GalNAc-type O-glycans: O-linked glycans are attached to serine or threonine (Ser/Thr) on peptides, where GalNAc is the root carbohydrate. This is mediated by a family of about 20 Golgi-resident polypeptide N-acetylgalactosaminyltransferases (ppGalNAcTs or GALNTs). Core 1 structures result from the attachment of β1-3 linked galactose to the core GalNAc using C1GALT1 and its chaperone C1GALT1C1. Core 2 structures then form upon addition of β1-6 linked GlcNAc by GCNT1. In addition to the GalNAc, Gal, and GlcNAc transferases, sialyltransferases such as ST6GALNAC1, 2 and ST3GAL1 are included as these mediate common O-linked sialylation modifications. Also included are the core-1 extension enzyme B3GalNT3 and the O-glycan specific sulfotransferase CHST4. Finally, modifications of core-3 and core-4 glycans can occur during disease and thus this ontology includes core-3 forming B3GNT6 and core-4 forming GCNT3. Other O-glycan core-types are rare in nature.

8. Chondroitin Sulfate & Heparan Sulfate Initiation: Chondroitin and heparan sulfate glycosaminoglycans all have a common core carbohydrate sequence attaching them to their proteins. These are constructed by the activity of specific Xylotransferases (XYLT1, XYLT2), galactosyltransferses B4GALT7 and B3GALT6 that sequentially add two galactose residues to Xylose, and the Glucuronyltransferase B3GAT3 then adds glucuronic acid to the terminal galactose. Also involved in the formation of this core is FAM20B, a kinase that 2-O-

phosphorylates Xylose. At this point, the addition of GalNAc to GlcA by CSGALNACT1 & 2 results in the initiation of chondroitin sulfates chains. The attachment of GlcNAc by EXTL3 to the same GlcA results in heparan sulfates.

9. Chondroitin/dermatan sulfate extension: Chondroitin sulfates and dermatan sulfates are extended via the addition of GalNAc-GlcA repeat units. This is catalyzed by CSGALNACT1 which is better suited for the initial GalNAc attachment followed by CSGALNACT2 which is preferred for synthesizing disaccharide repeats. CHSY1, CHSY3, CHPF and CHPF2, all exhibit dual β1,3GlcAT and β1,4GlcAT activity. Additional enzymes mediate sulfation. Epimerization of glucuronic acid to iduronic acid by DSE and DSEL results in the conversion of chondroitin sulfates to dermatan sulfates.

10. Heparan sulfate extension: EXT1 and EXT2 both have GlcUA and GlcNAc transferase activities and are together responsible for HS chain polymerization. EXTL1-3 are additional enzymes with GlcNAc transferase activity that facilitate heparin sulfate biosynthesis. Additional enzymes that are critical for heparin sulfate function include the HS2/3/6ST sulfotransferases, the GlcA epimerase GLCE and additional enzymes mediating N-sulfation (NDSTs).

11. Hyaluronan Synthesis: This pathway consists of the three hyaluronan synthases HAS1-3.

12. GPI Anchor Extension: This pathway includes glycogenes responsible for the synthesis of glycosphosphatidylinositol (GPI) anchored proteins in the ER. This involves synthesis of a glycan-lipid precursor that is en bloc transferred to proteins.

13. O-Mannose: This is initiated by the addition of mannose to Ser/Thr using POMT1 or POMT2. β1-2 or β1-4 GlcNAc linkages can then be made using POMGNT1 or POMGNT2 to yield M1 or M3 O-linked mannose structures, respectively. MGAT5B can facilitate β1-4 GlcNAc linkage onto the M1 structure to yield the M2 core. Additional carbohydrates typically found on complex N-linked glycan antennae can then attached. In particular, such extensions may be initiated by members of the B4GALT family or B3GALNT2. Specific variants are noted on α-dystroglycans.

14. O-linked Fucose: This pathway includes POFUT1, the enzyme responsible for the addition of fucose to Ser/Thr residues. MFNG, LFNG and RFNG which can attach β3GlcNAc to this fucose. B4GALT enzymes are included to account for galactose addition to this GlcNAc, and α2-3 or α2-6 sialyltransferases (ST3GAL or ST6GAL) are included as well as these can be terminal modifications.

15. Type 1 & 2 LacNAc: These enzymes help construct either Galβ1,3GlcNAc (Type 1) or Galβ1,4GlcNAc (Type 2) lactosamine chains on antennae of N-linked glycan, O-linked glycans and glycolipids. Also included are GCNT1-4 that can facilitate formation of I-branches on N-glycans.

16. Sialylation: This group encompasses all kinds of sialyltransferases: ST6GAL, ST3GAL, ST8SIA, and ST6GALNACs. Enrichments to this pathway capture overall increase in sialylation regardless of context.

17. Fucosylation: These include α1-2 (FUT1, 2) and α1-3 (FUT3, 4, 5, 6, 7, 9) fucosyltransferases that can act on N-glycans, O-glycans and glycolipids.

18. Sulfated glycan epitopes: This includes the enzymes forming the HNK1 epitope (B3GAT1, B3GAT2, CHST10) and sulfated sialyl Lewis-X structures.

19. ABO blood Group Synthesis: These are enzymes involved in the biosynthesis of ABO antigens

20. LacDiNAc: Glycogenes involved in the synthesis of LacDiNac and sulfated LacDiNac structures.

Establishing transcription factor–glycogene relationship: ChIP-Seq data from cancer cell lines and gene expression correlations from the TCGA data were downloaded from the Cistrome Cancer website for 29 cancer types in tab-limited form (http://cistrome.org/CistromeCancer/CancerTarget/)[15] . The data include the following fields: TF name, target gene, regulatory potential (RP) of TF to gene relationships, and Spearman's

correlation (ρ) between the TF and gene. Data from all 29 cancer types were agglomerated into one table, with an additional column specifying the cancer type for individual entries. The data were filtered for the 341 glycogenes in this manuscript (**Supplemental Table S1**). In total, the full dataset contained 41,771 TF-to-glycogene relationships, including relational data for 568 unique TFs found in the 29 cancer systems across all the glycogenes. Strong relationships between TFs and glycogenes were selected based on RP ≥ 0.5 and ρ ≥ 0.4 (Figure 2). This filtering resulted in 20,617 TF-glycogene relationships including 524 unique TFs across 29 cancer types.

Cytoscape was used to visualize TF-glycogene regulatory relationships[56] . To achieve this, all TF-glycogene relationship data were loaded into cytoscape as a network. These data were filtered based on RP and ρ thresholds defined previously. A binding potential (BP) score was computed by taking the product of RP and ρ for each TF-glycogene relationship. TF-glycogene relationships for each cancer type were separated into sub-networks. The prefuse force directed layout algorithm in cytoscape was used to arrange nodes in each cancer sub-network. The closeness of nodes to one another is weighted by 1-BP. Thus, nodes with high BPs will be placed closer together, whereas smaller BPs will be placed further away. Communities of glycogenes were detected using the clusterMaker feature of Cytoscape[17] . TF-glycogene interactions in each community were subjected to Reactome overrepresentation analyses to identify enriched signaling and glycosylation pathways.

Relating TF-glycogene interaction to glycosylation and signaling pathways: A Fisher's Exact Test was applied to determine if particular TF disproportionately regulate the 20 glycosylation pathways described in **Supplemental Table S3**. To achieve this, a contingency table was generated for each TF interaction with glycogenes present in each glycosylation pathway. This table included: i. Field A: The number of times the TF of interest interacted with a glycogene found **IN** the glycosylation pathway of interest. ii. Field B: The number of times the TF

of interest interacted with a glycogene **NOT IN** the pathway of interest. iii. Field C: The number of times other TFs **NOT** of interest interacted with a glycogene **IN** the glycosylation pathway of interest. iv. Field D: The number of times others TFs **NOT** of interest regulated glycogenes **NOT IN** the pathway of interest. The total number of contingency tables generated was thus: TF × glycosylation pathways × cancer types. Fisher's exact test $p \leq 0.05$ was used to determine statistically significant TFs enriched in each glycosylation pathway.

The Reactome DB was used as a reference to associate the TF-pathway associations above with cell signaling. Here, TFs enriched in each glycosylation pathway were submitted to the Reactome's over-representation analysis API to associate the TFs with signaling pathways[14] . Pathway enrichments with adjusted p (FDR)<0.1 were considered to be statistically significant. The connection between cell signaling pathways and TFs, and that between the TFs and glycosylation pathways were visualized using alluvial plots generated using the R package ggalluvial. Only signaling pathways with < 30 members are presented, as they may be more specific functional regulators of glycosylation.

Robust Rank Aggregation Analysis: Robust Rank Aggregation (RRA) was performed using the R package RobustRankAggreg[57] . Here, TF-glycogene relations were sorted in descending order based on RP values for each of the glycogenes present in the 20 glycosylation pathways, individually. They were then ranked based on this operation. The ranks were then normalized based on the number of total TFs associated with all the glycogenes in that pathway. This ranked list was independently generated for each of the 29 cancer types, and used as input for the "aggregateRanks" function of the RobustRankAggreg package. The function computes the likelihood using the binomial distribution expression. TFs with RRA p-values ≤ 0.1 were considered to be statistically significant, and were considered to pervasively regulate a glycosylation pathway across cancer types.

SUPPORTING INFORMATION

Supplementary Table S1:
File Name: TableS1_Glycogenes.xlsx
File Format: XLXS
Title: List of 341 glycogenes used for cystoscape maps

Supplementary Table S2:
File Name: TableS2_CancerTypes.xlsx
File Format: XLSX
Title: Cancer type list

Supplementary Table S3:
File Name: TableS1_Glycogene_Pathway_Lists.xlsx
File Format: XLSX
Title: Glycogene pathway lists

Supplementary Table S4:
File Name: TableS4_Fishers_exact_test_summary.xlsx
File Format: XLXS
Title: Fisher's exact test to infer TF-glycosylation pathway relation ($p<0.05$ data are highlighted)

Supplementary Table S5:
File Name: TableS5_Reactome_Enrich_Pathways.xlsx
File Format: XLXS
Title: Reactome pathway enrichments for all TFs

Supplementary Table S6:
File Name: TableS6_Robust_Rank_aggregation.xlsx
File Format: XLXS
Title: RRA results showing TFs that more commonly regulate glycogenes in given pathway, across cancer types: ($p<0.1$ are highlighted)

Supplementary File S1:
File Name: FileS1_CancerNetworks_Cistrome.cys
File Format : cys (Cytoscape Session File)
Title: Cistrome Cancer TF-to-glycogene subnetworks

Supplementary File S2:
File Name : FileS2_supplemental_alluvials.pdf
File Format: PDF
Title: Alluvial plots for all cancer types

ACKNOWLEDGEMENT

We gratefully acknowledge helpful discussions with Prof. Rudiyanto Gunawan

FUNDING

This work was supported US National Institutes of Health grants HL103411, GM133195 and GM126537.

REFERENCES:

(1) Zhu, Y.; Groth, T.; Kelkar, A.; Zhou, Y.; Neelamegham, S. A GlycoGene CRISPR-Cas9 Lentiviral Library to Study Lectin Binding and Human Glycan Biosynthesis Pathways. *Glycobiology* **2020**. https://doi.org/10.1093/glycob/cwaa074.

(2) Lambert, S. A.; Jolma, A.; Campitelli, L. F.; Das, P. K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T. R.; Weirauch, M. T. The Human Transcription Factors. *Cell* **2018**, *172* (4), 650–665. https://doi.org/10.1016/j.cell.2018.01.029.

(3) Namba, S.; Sato, K.; Kojima, S.; Ueno, T.; Yamamoto, Y.; Tanaka, Y.; Inoue, S.; Nagae, G.; Iinuma, H.; Hazama, S.; Ishihara, S.; Aburatani, H.; Mano, H.; Kawazu, M. Differential Regulation of CpG Island Methylation within Divergent and Unidirectional Promoters in Colorectal Cancer. *Cancer Sci.* **2019**, *110* (3), 1096–1104. https://doi.org/10.1111/cas.13937.

(4) Malta, T. M.; de Souza, C. F.; Sabedot, T. S.; Silva, T. C.; Mosella, M. S.; Kalkanis, S. N.; Snyder, J.; Castro, A. V. B.; Noushmehr, H. Glioma CpG Island Methylator Phenotype (G-CIMP): Biological and Clinical Implications. *Neuro. Oncol.* **2018**, *20* (5), 608–620. https://doi.org/10.1093/neuonc/nox183.

(5) Wolters-Eisfeld, G.; Mercanoglu, B.; Strohmaier, A.; Guengoer, C.; Izbicki, J. R.; Bockhorn, M. Hypoxia Induced HIF1a-Mediated O-GalNAc Glycosylation of Cytosolic O-GlcNAcylated Proteins to Regulate Signaling Pathways in Pancreatic Cancer. *J. Clin. Oncol.* **2017**, *35* (15_suppl), e15739–e15739. https://doi.org/10.1200/JCO.2017.35.15_suppl.e15739.

(6) Tu, C.-F.; Wu, M.-Y.; Lin, Y.-C.; Kannagi, R.; Yang, R.-B. FUT8 Promotes Breast Cancer Cell Invasiveness by Remodeling TGF-β Receptor Core Fucosylation. *Breast Cancer Res.* **2017**, *19* (1), 111. https://doi.org/10.1186/s13058-017-0904-8.

(7) Neelamegham, S.; Mahal, L. K. Multi-Level Regulation of Cellular Glycosylation: From Genes to Transcript to Enzymes to Structure. *Curr. Opin. Struct. Biol.* **2016**, *21* (2), 145–152. https://doi.org/10.5588/ijtld.16.0716.Isoniazid.

(8) Furey, T. S. ChIP-Seq and beyond: New and Improved Methodologies to Detect and Characterize Protein-DNA Interactions. *Nat. Rev. Genet.* **2012**, *13* (12), 840–852. https://doi.org/10.1038/nrg3306.

(9) Truax, A. D.; Greer, S. F. ChIP and Re-ChIP Assays: Investigating Interactions between Regulatory Proteins, Histone Modifications, and the DNA Sequences to Which They Bind. *Methods Mol. Biol.* **2012**, *809*, 175–188. https://doi.org/10.1007/978-1-61779-376-9_12.

(10) Desvoyes, B.; Sequeira-Mendes, J.; Vergara, Z.; Madeira, S.; Gutierrez, C. Sequential ChIP Protocol for Profiling Bivalent Epigenetic Modifications (ReChIP). *Methods Mol. Biol.* **2018**, *1675*, 83–97. https://doi.org/10.1007/978-1-4939-7318-7_6.

(11) Bayat, P.; Nosrati, R.; Alibolandi, M.; Rafatpanah, H.; Abnous, K.; Khedri, M.; Ramezani, M. SELEX Methods on the Road to Protein Targeting with Nucleic Acid Aptamers. *Biochimie* **2018**, *154*, 132–155. https://doi.org/10.1016/j.biochi.2018.09.001.

(12) Zhao, Y.; Granas, D.; Stormo, G. D. Inferring Binding Energies from Selected Binding Sites. *PLoS Comput. Biol.* **2009**, *5* (12), e1000590. https://doi.org/10.1371/journal.pcbi.1000590.

(13) Grossman, R. L.; Heath, A. P.; Ferretti, V.; Varmus, H. E.; Lowy, D. R.; Kibbe, W. A.; Staudt, L. M. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **2016**, *375* (12), 1109–1112. https://doi.org/10.1056/NEJMp1607591.

(14) Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; Loney, F.; May, B.; Milacic, M.; Rothfels, K.; Sevilla, C.; Shamovsky, V.; Shorser, S.; Varusai, T.; Weiser, J.; Wu, G.; Stein, L.; Hermjakob, H.; D'Eustachio, P. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **2020**, *48* (D1), D498–D503. https://doi.org/10.1093/nar/gkz1031.

(15) Mei, S.; Meyer, C. A.; Zheng, R.; Qin, Q.; Wu, Q.; Jiang, P.; Li, B.; Shi, X.; Wang, B.; Fan, J.; Shih, C.; Brown, M.; Zang, C.; Liu, X. S. Cistrome Cancer: A Web Resource for Integrative Gene Regulation Modeling in Cancer. *Cancer Res.* **2017**, *77* (21), e19–e22. https://doi.org/10.1158/0008-5472.CAN-17-0327.

(16) Bernard, P. S.; Parker, J. S.; Mullins, M.; Cheung, M. C. U.; Leung, S.; Voduc, D.; Vickery, T.; Davies, S.; Fauron, C.; He, X.; Hu, Z.; Quackenhush, J. F.; Stijleman, I. J.; Palazzo, J.; Matron, J. S.; Nobel, A. B.; Mardis, E.; Nielsen, T. O.; Ellis, M. J.; Perou, C. M. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **2009**, *27* (8), 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370.

(17) Morris, J. H.; Apeltsin, L.; Newman, A. M.; Baumbach, J.; Wittkop, T.; Su, G.; Bader, G. D.; Ferrin, T. E. ClusterMaker: A Multi-Algorithm Clustering Plugin for Cytoscape. *BMC Bioinformatics* **2011**, *12* (1), 436. https://doi.org/10.1186/1471-2105-12-436.

(18) Miklossy, G.; Hilliard, T. S.; Turkson, J. Therapeutic Modulators of STAT Signalling for Human Diseases. *Nat. Rev. Drug Discov.* **2013**, *12* (8), 611–629. https://doi.org/10.1038/nrd4088.

(19) Bi, X.; Hameed, M.; Mirani, N.; Pimenta, E. M.; Anari, J.; Barnes, B. J. Loss of Interferon Regulatory Factor 5 (IRF5) Expression in Human Ductal Carcinoma Correlates with Disease Stage and Contributes to Metastasis. *Breast Cancer Res.* **2011**, *13* (6), R111. https://doi.org/10.1186/bcr3053.

(20) Yanai, H.; Negishi, H.; Taniguchi, T. The IRF Family of Transcription Factors: Inception, Impact and Implications in Oncogenesis. *Oncoimmunology* **2012**, *1* (8), 1376–1386. https://doi.org/10.4161/onci.22475.

(21) Sampieri, L.; Di Giusto, P.; Alvarez, C. CREB3 Transcription Factors: ER-Golgi Stress Transducers as Hubs for Cellular Homeostasis. *Front. Cell Dev. Biol.* **2019**, *7*, 123. https://doi.org/10.3389/fcell.2019.00123.

(22) Khan, H. A.; Margulies, C. E. The Role of Mammalian Creb3-Like Transcription Factors in Response to Nutrients. *Front. Genet.* **2019**, *10*, 591. https://doi.org/10.3389/fgene.2019.00591.

(23) Pu, Q.; Lu, L.; Dong, K.; Geng, W.; Lv, Y.; Gao, H. The Novel Transcription Factor CREB3L4 Contributes to the Progression of Human Breast Carcinoma. *J. Mammary Gland Biol. Neoplasia* **2020**, *25* (1), 37–50. https://doi.org/10.1007/s10911-020-09443-6.

(24) Dimitroff, C. J. I-Branched Carbohydrates as Emerging Effectors of Malignant Progression. *Proc. Natl. Acad. Sci.* **2019**, *116* (28), 13729–13737. https://doi.org/10.1073/pnas.1900268116.

(25) Romagnoli, M.; Belguise, K.; Yu, Z.; Wang, X.; Landesman-Bollag, E.; Seldin, D. C.; Chalbos, D.; Barillé-Nion, S.; Jézéquel, P.; Seldin, M. L.; Sonenshein, G. E. Epithelial-to-Mesenchymal Transition Induced by TGF-B1 Is Mediated by Blimp-1-Dependent Repression of BMP-5. *Cancer Res.* **2012**, *72* (23), 6268–6278. https://doi.org/10.1158/0008-5472.CAN-12-2270.

(26) Denechaud, P.-D.; Fajas, L.; Giralt, A. E2F1, a Novel Regulator of Metabolism. *Front. Endocrinol. (Lausanne).* **2017**, *8*, 311. https://doi.org/10.3389/fendo.2017.00311.

(27) Hollern, D. P.; Swiatnicki, M. R.; Rennhack, J. P.; Misek, S. A.; Matson, B. C.; McAuliff, A.; Gallo, K. A.; Caron, K. M.; Andrechek, E. R. E2F1 Drives Breast Cancer Metastasis by Regulating the Target Gene FGF13 and Altering Cell Migration. *Sci. Rep.* **2019**, *9* (1), 10718. https://doi.org/10.1038/s41598-019-47218-0.

(28) Musa, J.; Aynaud, M.-M.; Mirabeau, O.; Delattre, O.; Grünewald, T. G. MYBL2 (B-Myb): A Central Regulator of Cell Proliferation, Cell Survival and Differentiation Involved in Tumorigenesis. *Cell Death Dis.* **2017**, *8* (6), e2895. https://doi.org/10.1038/cddis.2017.244.

(29) Pon, J. R.; Marra, M. A. MEF2 Transcription Factors: Developmental Regulators and Emerging Cancer Genes. *Oncotarget* **2016**, *7* (3), 2297–2312. https://doi.org/10.18632/oncotarget.6223.

(30) Yu, W.; Huang, C.; Wang, Q.; Huang, T.; Ding, Y.; Ma, C.; Ma, H.; Chen, W. MEF2 Transcription Factors Promotes EMT and Invasiveness of Hepatocellular Carcinoma through TGF-B1 Autoregulation Circuitry. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **2014**, *35* (11), 10943–10951. https://doi.org/10.1007/s13277-014-2403-1.

(31) Nan, X.; Ng, H. H.; Johnson, C. A.; Laherty, C. D.; Turner, B. M.; Eisenman, R. N.; Bird, A. Transcriptional Repression by the Methyl-CpG-Binding Protein MeCP2 Involves a Histone Deacetylase Complex. *Nature* **1998**, *393* (6683), 386–389. https://doi.org/10.1038/30764.

(32) Jones, P. L.; Veenstra, G. J.; Wade, P. A.; Vermaak, D.; Kass, S. U.; Landsberger, N.; Strouboulis, J.; Wolffe, A. P. Methylated DNA and MeCP2 Recruit Histone Deacetylase to Repress Transcription. *Nat. Genet.* **1998**, *19* (2), 187–191. https://doi.org/10.1038/561.

(33) Tong, D.; Zhang, J.; Wang, X.; Li, Q.; Liu, L. Y.; Yang, J.; Guo, B.; Ni, L.; Zhao, L.; Huang, C. MeCP2 Facilitates Breast Cancer Growth via Promoting Ubiquitination-Mediated P53 Degradation by Inhibiting RPL5/RPL11 Transcription. *Oncogenesis* **2020**, *9* (5), 56. https://doi.org/10.1038/s41389-020-0239-7.

(34) Sharma, K.; Singh, J.; Frost, E. E.; Pillai, P. P. MeCP2 Overexpression Inhibits Proliferation, Migration and Invasion of C6 Glioma by Modulating ERK Signaling and Gene Expression. *Neurosci. Lett.* **2018**, *674*, 42–48. https://doi.org/10.1016/j.neulet.2018.03.020.

(35) Massagué, J. TGFβ Signalling in Context. *Nat. Rev. Mol. Cell Biol.* **2012**, *13* (10), 616–630. https://doi.org/10.1038/nrm3434.

(36) Deckers, M.; van Dinther, M.; Buijs, J.; Que, I.; Löwik, C.; van der Pluijm, G.; ten Dijke, P. The Tumor Suppressor Smad4 Is Required for Transforming Growth Factor Beta-Induced Epithelial to Mesenchymal Transition and Bone Metastasis of Breast Cancer Cells. *Cancer Res.* **2006**, *66* (4), 2202–2209. https://doi.org/10.1158/0008-5472.CAN-05-3560.

(37) Liu, N.; Xi, Y.; Callaghan, M. U.; Fribley, A.; Moore-Smith, L.; Zimmerman, J. W.; Pasche, B.; Zeng, Q.; Li, Y. SMAD4 Is a Potential Prognostic Marker in Human Breast Carcinomas. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **2014**, *35* (1), 641–650. https://doi.org/10.1007/s13277-013-1088-1.

(38) MacDonald, B. T.; Tamai, K.; He, X. Wnt/Beta-Catenin Signaling: Components, Mechanisms, and Diseases. *Dev. Cell* **2009**, *17* (1), 9–26. https://doi.org/10.1016/j.devcel.2009.06.016.

(39) Ip, W.; Chiang, Y.-T. A.; Jin, T. The Involvement of the Wnt Signaling Pathway and TCF7L2 in Diabetes Mellitus: The Current Understanding, Dispute, and Perspective. *Cell Biosci.* **2012**, *2* (1), 28. https://doi.org/10.1186/2045-3701-2-28.

(40) Wenzel, J.; Rose, K.; Haghighi, E. B.; Lamprecht, C.; Rauen, G.; Freihen, V.; Kesselring, R.; Boerries, M.; Hecht, A. Loss of the Nuclear Wnt Pathway Effector TCF7L2 Promotes Migration and Invasion of Human Colorectal Cancer Cells. *Oncogene* **2020**, *39* (19), 3893–3909. https://doi.org/10.1038/s41388-020-1259-7.

(41) Chen, F.; Liu, X.; Bai, J.; Pei, D.; Zheng, J. The Emerging Role of RUNX3 in Cancer Metastasis (Review). *Oncol Rep* **2016**, *35* (3), 1227–1236. https://doi.org/10.3892/or.2015.4515.

(42) Tian, F.; DaCosta Byfield, S.; Parks, W. T.; Yoo, S.; Felici, A.; Tang, B.; Piek, E.; Wakefield, L. M.; Roberts, A. B. Reduction in Smad2/3 Signaling Enhances Tumorigenesis but Suppresses Metastasis of Breast Cancer Cell Lines. *Cancer Res.* **2003**, *63* (23), 8284–8292.

(43) Petersen, M.; Pardali, E.; van der Horst, G.; Cheung, H.; van den Hoogen, C.; van der Pluijm, G.; Ten Dijke, P. Smad2 and Smad3 Have Opposing Roles in Breast Cancer Bone Metastasis by Differentially Affecting Tumor Angiogenesis. *Oncogene* **2010**, *29* (9), 1351–1361. https://doi.org/10.1038/onc.2009.426.

(44) Oh, S.; Shin, S.; Janknecht, R. The Small Members of the JMJD Protein Family: Enzymatic Jewels or Jinxes? *Biochim. Biophys. Acta - Rev. Cancer* **2019**, *1871* (2), 406–418. https://doi.org/https://doi.org/10.1016/j.bbcan.2019.04.002.

(45) Chen, C.; Aihemaiti, M.; Zhang, X.; Qu, H.; Sun, Q.-L.; He, Q.-S.; Yu, W.-B. Downregulation of Histone Demethylase JMJD1C Inhibits Colorectal Cancer Metastasis through Targeting ATF2. *Am. J. Cancer Res.* **2018**, *8* (5), 852–865.

(46) Klümper, N.; Syring, I.; Vogel, W.; Schmidt, D.; Müller, S. C.; Ellinger, J.; Shaikhibrahim, Z.; Brägelmann, J.; Perner, S. Mediator Complex Subunit MED1 Protein Expression Is Decreased during Bladder Cancer Progression. *Front. Med.* **2017**, *4*, 30. https://doi.org/10.3389/fmed.2017.00030.

(47) Leonard, M.; Zhang, X. Estrogen Receptor Coactivator Mediator Subunit 1 (MED1) as a Tissue-Specific Therapeutic Target in Breast Cancer. *J. Zhejiang Univ. Sci. B* **2019**, *20* (5), 381–390. https://doi.org/10.1631/jzus.B1900163.

(48) Liu, H.; Dai, X.; Cao, X.; Yan, H.; Ji, X.; Zhang, H.; Shen, S.; Si, Y.; Zhang, H.; Chen, J.; Li, L.; Zhao, J. C.; Yu, J.; Feng, X.-H.; Zhao, B. PRDM4 Mediates YAP-Induced Cell Invasion by Activating Leukocyte-Specific Integrin B2 Expression. *EMBO Rep.* **2018**, *19* (6). https://doi.org/10.15252/embr.201745180.

(49) San-Marina, S.; Han, Y.; Suarez Saiz, F.; Trus, M. R.; Minden, M. D. Lyl1 Interacts with CREB1 and Alters Expression of CREB1 Target Genes. *Biochim. Biophys. Acta - Mol. Cell Res.* **2008**, *1783* (3), 503–517. https://doi.org/https://doi.org/10.1016/j.bbamcr.2007.11.015.

(50) Lin, S.; Kemmner, W.; Grigull, S.; Schlag, P. M. Cell Surface A2, 6-Sialylation Affects Adhesion of Breast Carcinoma Cells. *Exp. Cell Res.* **2002**, *276* (1), 101–110. https://doi.org/10.1006/excr.2002.5521.

(51) Cui, H.; Lin, Y.; Yue, L.; Zhao, X.; Liu, J. Differential Expression of the A2,3-Sialic Acid Residues in Breast Cancer Is Associated with Metastatic Potential. *Oncol. Rep.* **2011**, *25* (5), 1365–1371. https://doi.org/10.3892/or.2011.1192.

(52) Yevshin, I.; Sharipov, R.; Kolmykov, S.; Kondrakhin, Y.; Kolpakov, F. GTRD: A Database on Gene Transcription Regulation-2019 Update. *Nucleic Acids Res.* **2019**, *47* (D1), D100–D105. https://doi.org/10.1093/nar/gky1128.

(53) Marbach, D.; Lamparter, D.; Quon, G.; Kellis, M.; Kutalik, Z.; Bergmann, S. Tissue-Specific Regulatory Circuits Reveal Variable Modular Perturbations across Complex Diseases. *Nat. Methods* **2016**, *13* (4), 366–370. https://doi.org/10.1038/nmeth.3799.

(54) Taniguchi, N.; Honke, K.; Fukuda, M.; Narimatsu, H.; Yamaguchi, Y.; Angata, T. *Handbook of Glycosyltransferases and Related Genes, Second Edition*, 2nd ed.; Springer, 2014; Vol. 1–2. https://doi.org/10.1007/978-4-431-54240-7.

(55) Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Darvill, A. G.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H. *Essentials of Glycobiology*.

(56) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504. https://doi.org/10.1101/gr.1239303.

(57) Kolde, R.; Laur, S.; Adler, P.; Vilo, J. Robust Rank Aggregation for Gene List Integration and Meta-Analysis. *Bioinformatics* **2012**. https://doi.org/10.1093/bioinformatics/btr709.

Groth & Neelamegham
Figure 1

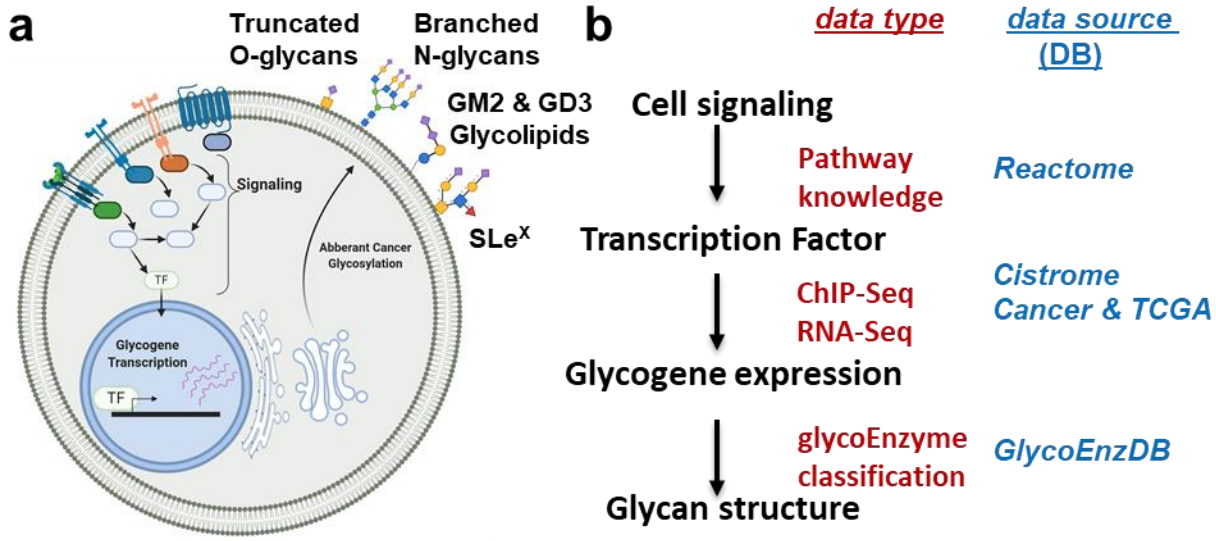

Figure 1. A systems glycobiology framework to link multi-OMICs data: **a.** Cell signaling proceeds to trigger transcription factor (TF) activity. The binding of TFs to sites proximal to the transcriptional start site triggers glycogene expression. A complex set of reaction pathways then results in the synthesis of various carbohydrate types, many of which are either secreted or expressed on the cell surface. **b.** Data available at various resources can establish the link between cell signaling and glycan biosynthesis. The Reactome DB contains vast cell signaling knowledge. Chip-Seq and RNA-Seq data available at the Cistrome Cancer DB describe the link between the TFs and glycogenes. Pathway curation at the GlycoEnzDB establishes the link between glycogenes and glycan structures. Cell illustration created using BioRender.com.

Groth & Neelamegham
Figure 2

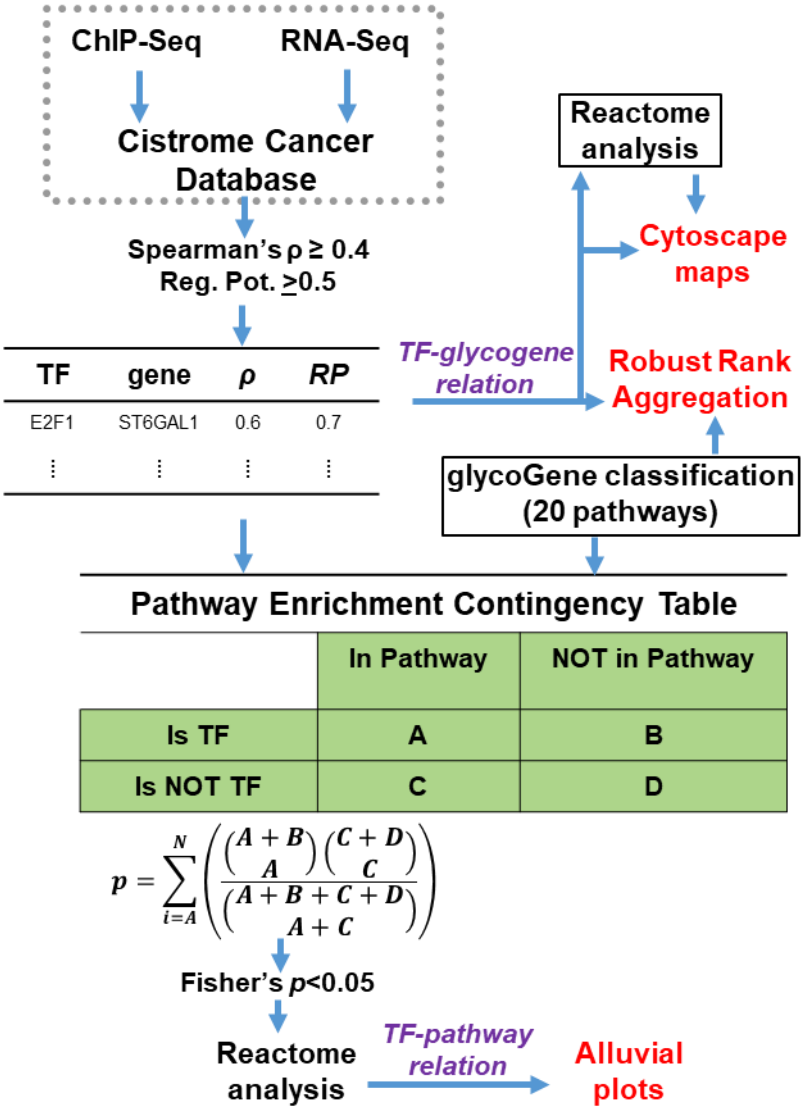

Figure 2. Analysis workflow to establish TF-glycogene, and TF-glycosylation pathway relationships: ChiP-seq provides evidence of TF binding to promoter regions with regulatory potential ($0 \leq RP \leq 1$) quantifying the likelihood that this is functionally important. RNA-Seq quantifies Spearman's correlation (ρ) between TF and gene expression. Filtering these based on data available at the Cistrome Cancer DB establishes potential TF-glycogene interactions in specific cancers. Cytoscape maps relating TFs to glycogenes and ReactomeDB signaling pathways was established, These data are also used for RRA analysis. Whether a candidate TFs significantly and specifically regulates any of the 20 manually curated glycosylation pathways was determined by developing contingency tables for each TF-glycosylation pathway interactions, and analyzing using the Fisher's exact test. Here, 'A' counts the number of TF-glycogene interactions in the glycosylation pathway of interest (i.e. A=count[(t=TF)&(g∈G)]). Here, TF & G = Transcription factor and glycogenes in specific pathway being tested for enrichment; t & g = Highly correlated TF-glycogene pairs that are being tested. Similarly, B=count[(t=TF)&(g∉G)]; C=count[(t≠TF)&(g∈G)]; D=count[(t≠TF)&(g∉G)]. 'N' is the number of candidate genes in the pathway. ReactomeDB analysis was performed for these selected TFs. Alluvial plots displayed the relation between cell signaling-TF-glycosylation pathways.

Groth & Neelamegham
Figure 3

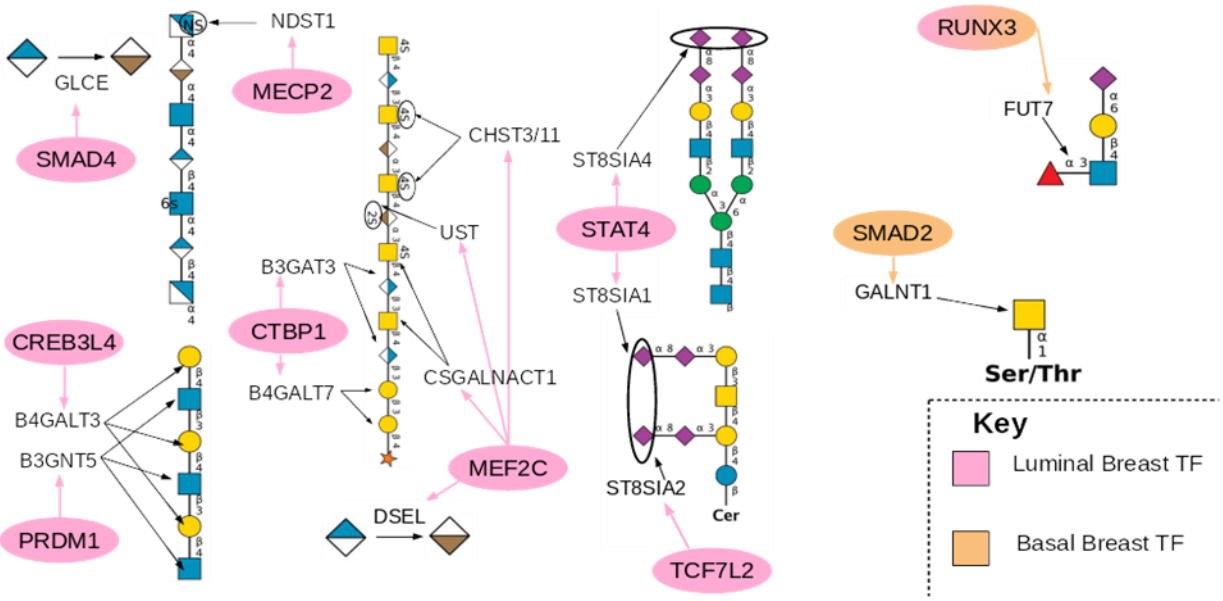

Figure 3. Summary of all TFs enriched to glycosylation pathways for luminal and basal breast cancer: The TFs found to be enriched to glycogenes are shown in pink for luminal and orange for basal breast cancer. The glycans synthesized by the enriched glycogenes are shown in SNFG format (https://www.ncbi.nlm.nih.gov/glycans/snfg.html).

Groth & Neelamegham
Figure 4

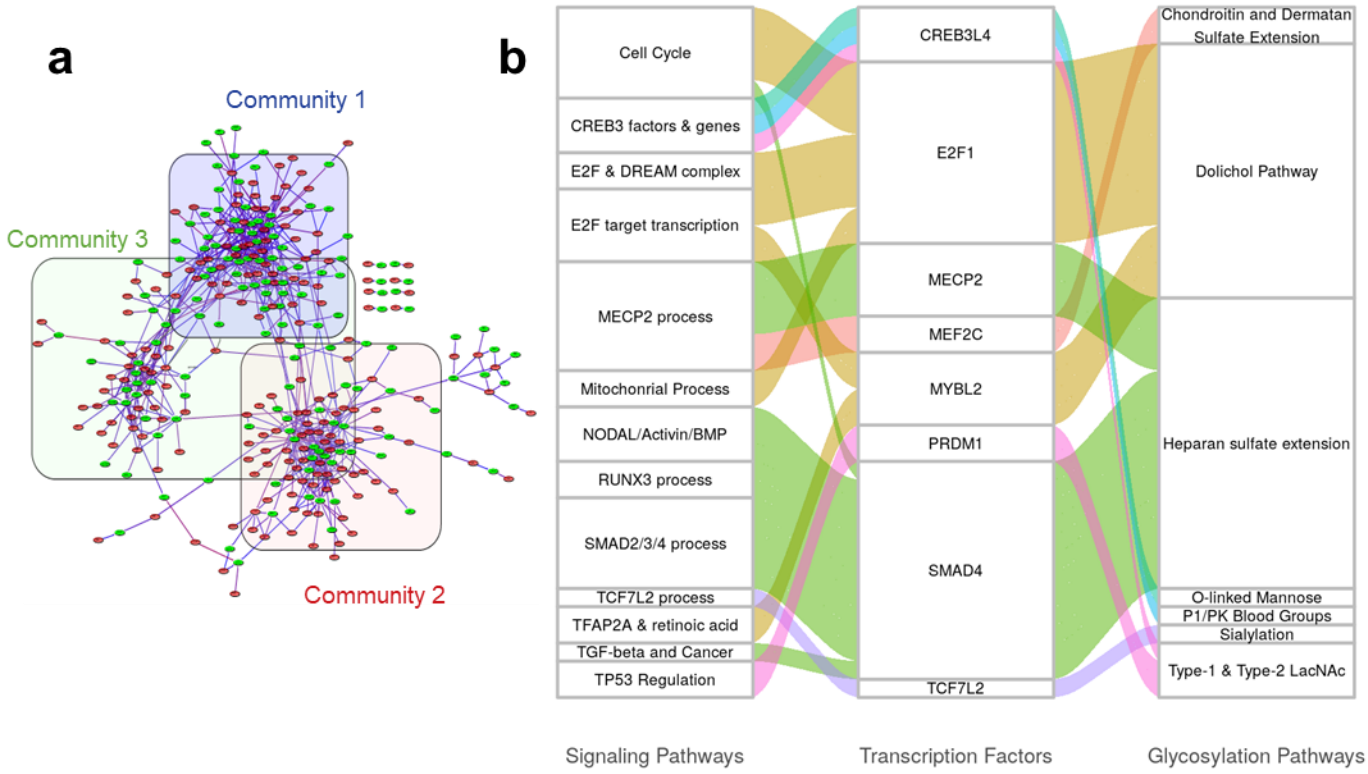

Figure 4. Luminal breast cancer signaling pathway enrichment and glycogene connections: a. *TF-to-glycogene communities in luminal breast cancer*: Three large TF-to-glycogene communities were discovered in the luminal breast subnetwork. Community 1 was enriched for pathways involving RUNX3, RUNX1, IL-21, and PTEN, whereas communities 2 and 3 consist primarily of chromatin modifying enzymes. **b.** *Signaling pathway enrichment analysis for luminal breast cancer*: Connections between signaling pathways and transcription factors found to be statistically significant for luminal breast cancer. Some pathways enriched to TFs were condensed to conserve space. More TF-to-glycogene relationships exist in luminal breast cancer and these can be viewed in the cytoscape figures (**Supplemental Figure S1**).

Groth & Neelamegham
Figure 5

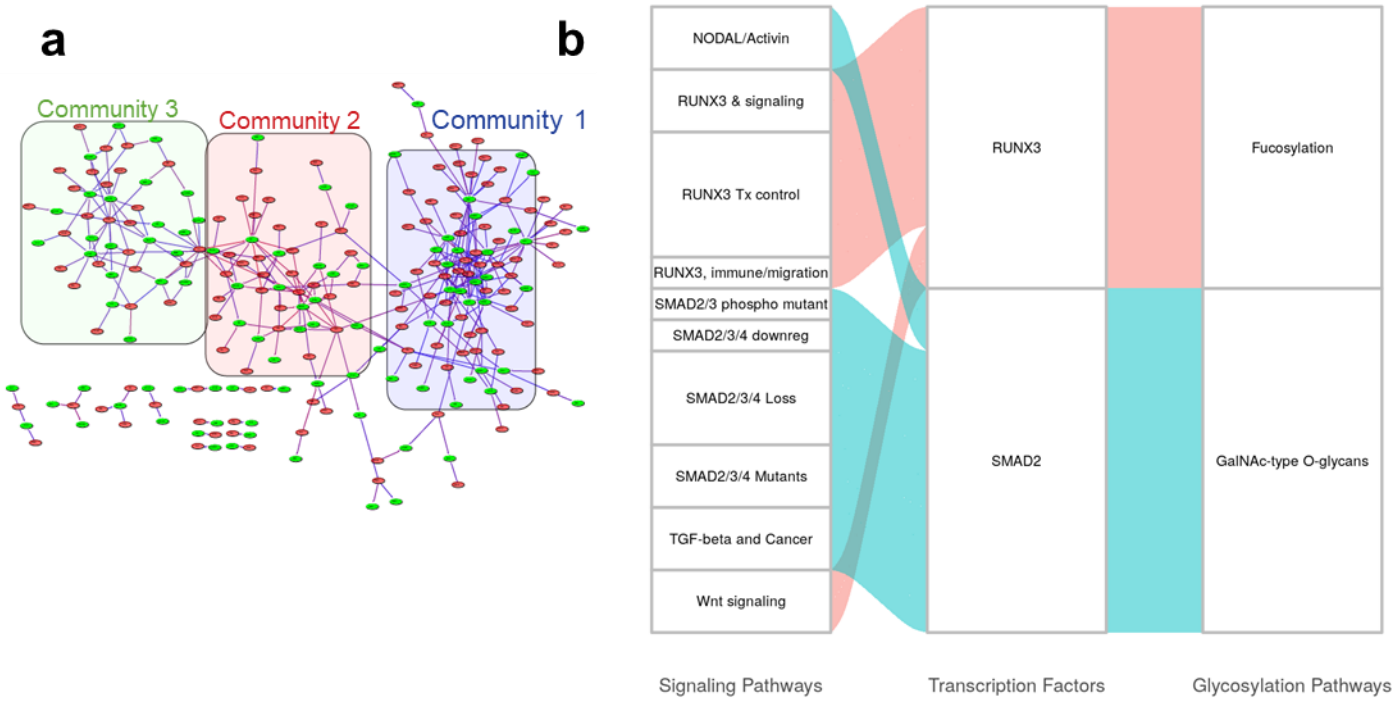

Figure 5. Basal breast cancer signaling pathway enrichments and glycogene connections: a. *TF-to-glycogene communities in basal breast cancer*: Three large TF-to-glycogene communities were discovered in the basal breast subnetwork. Community 1 has TFs enriched to chromatin modifying enzymes, and community 2 has TFs enriched to interferon α/β/γ signaling. Community 3 did not have any signaling pathways enriched. **b.** *Signaling pathway enrichment analysis for basal breast cancer*: Connections between signaling pathways and TFs found to be statistically significant for basal breast cancer. TFs displayed have been enriched to the displayed glycosylation pathways using the Fisher's exact test.