

Embryonic EMT programs establish proteoglycan expression patterns that persist in adult human tissues

Saikat Das Sajib

UiT The Arctic University of Norway

Anne Kristin McLaren Berge

UiT The Arctic University of Norway

Taufeeq Nii Korquaye Quaye

UiT The Arctic University of Norway

Kjersti Sellæg

UiT The Arctic University of Norway

Beate Hegge

UiT The Arctic University of Norway

Catherine L. R. Merry

University of Nottingham

Svein Olav Kolset

University of Oslo

Susannah von Hofsten

UiT The Arctic University of Norway

Erik Knutsen

erik.knutsen@uit.no

UiT The Arctic University of Norway

Research Article

Keywords: Proteoglycan core proteins, glycosaminoglycan biosynthetic enzymes, epithelial-mesenchymal transition transcription factors, embryonic germ layers, single-cell RNA sequencing

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11070727/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Proteoglycans and glycosaminoglycans (GAGs) are essential for embryonic development and adult tissue homeostasis. Proteoglycan core proteins largely determine proteoglycan localization, whereas enzymes involved in GAG initiation, elongation, modification, and degradation regulate the composition of attached GAG chains. Although proteoglycans are known to participate in epithelial-mesenchymal transition (EMT), it remains unclear whether developmental EMT programs contribute to shaping proteoglycan expression patterns that persist into adulthood.

Results

Using single-cell RNA sequencing data from 81 adult human cell types, we analyzed the expression of 42 proteoglycan core proteins and 48 GAG biosynthetic enzymes. Proteoglycan core protein expression was strongly associated with embryonic germ-layer origin and EMT transcription factor (TF) activity, suggesting that adult tissues retain transcriptional features established during embryogenesis. In contrast, GAG biosynthetic enzymes displayed more cell type- and tissue-specific expression patterns, including enrichment of chondroitin/dermatan sulfate biosynthetic enzymes in mesoderm-derived immune cells and broad GAG biosynthetic enzyme expression in central nervous system cells. Analysis of human embryonic single-cell data during gastrulation demonstrated dynamic, lineage-specific regulation of both proteoglycan core proteins and GAG biosynthetic enzymes in parallel with EMT TF expression. Approximately half of the proteoglycan core proteins retained their embryonic germ-layer specificity in adult tissues, whereas GAG biosynthetic enzymes showed substantially lower conservation. Integration of embryonic and adult transcriptomic data with EMT TF perturbation datasets and ChIP-seq analyses identified the ZEB family as prominent candidate regulators, primarily through enhancer-associated regulatory elements.

Conclusions

Together, these findings suggest that adult proteoglycan core protein expression retains a transcriptional imprint of embryonic EMT programs, whereas GAG biosynthetic enzymes become increasingly governed by cell type- and tissue-specific regulatory mechanisms. Our study provides evidence that developmental origin is a major determinant of the adult proteoglycan landscape and offers a framework for understanding how embryonic transcriptional programs contribute to extracellular matrix organization in health and disease.

Background

Proteoglycans are heavily glycosylated proteins that are widely expressed across human tissues and play essential roles in tissue development, organization, homeostasis, and remodeling (1, 2). They consist of a protein core decorated with covalently attached glycosaminoglycan (GAG) chains and are found on the cell surface, within intracellular granules, and throughout the extracellular matrix (ECM) (2). The diversity of both the core proteins and their attached GAG chains generates substantial structural and functional heterogeneity, enabling proteoglycans to regulate numerous biological processes.

Through interactions with extracellular ligands (3–6), other proteoglycans (6, 7), and cell adhesion molecules (2, 8), proteoglycans function as molecular interfaces between cells and their surrounding microenvironment. As components of the basement membrane, they contribute to defining tissue compartmentalization and establishing cell polarity (6, 9, 10). Additionally, proteoglycans act as receptors, co-receptors, and ligands for various growth factors, cytokines, and morphogens that provide essential signals for tissue homeostasis and regeneration (11–16).

Previous studies suggest that proteoglycans play a dual role in epithelial-mesenchymal transition (EMT), both regulating and being regulated by the process (17). EMT is a reversible developmental program during which epithelial cells lose polarity and cell-cell adhesion while acquiring migratory and mesenchymal characteristics. The process is initiated by signaling pathways including TGF- β , Wnt, Notch, and receptor tyrosine kinases, culminating in activation of the EMT transcription factors (TFs) SNAI1/2, TWIST1/2, and ZEB1/2, which repress epithelial genes while inducing mesenchymal gene expression (18). Remodeling of the extracellular matrix accompanies these transcriptional changes and facilitates cell migration (19).

EMT is initiated during gastrulation in the developing embryo, enabling the formation of the three primary germ layers that ultimately give rise to every cell type in the adult organism. Proteoglycans contribute to these developmental events by stabilizing concentration gradients of signaling molecules and enabling signaling pathways that are critical for cell determination and tissue organization (20–23). For instance, the formation of the endoderm and mesoderm germ layers from the single layer of epithelial cells termed the epiblast, relies on proteoglycans to promote fibroblast growth factor signaling (24, 25). The remaining cells of the epiblast become the ectoderm. Studies in model organisms have also shown that proteoglycan-dependent activation of Wnt signaling is critical for cell movement during gastrulation (26, 27), and that the relocation of epiblast cells requires ECM rearrangement, highlighting a potential role for proteoglycans in the basement membrane underlying the epiblast (28, 29). Because all adult tissues originate from these embryonic lineages, transcriptional programs established during gastrulation may continue to influence tissue-specific gene expression long after embryogenesis has concluded. Whether such developmental programs contribute to the proteoglycan expression patterns observed in adult tissues, however, remains unknown.

Although proteoglycans have long been recognized as regulators of EMT, it remains unclear whether EMT-associated transcriptional programs also establish the proteoglycan expression patterns observed in adult tissues. Furthermore, it is unknown whether proteoglycan core proteins and the enzymes

responsible for GAG biosynthesis are governed by similar developmental regulatory mechanisms or become independently regulated following lineage specification. Here, we integrate adult and embryonic single-cell transcriptomic datasets with EMT TF perturbation experiments and ChIP-seq analyses to test the hypothesis that embryonic EMT programs establish proteoglycan expression signatures that persist into adulthood while GAG biosynthetic enzymes acquire increasingly tissue-specific regulation.

Results

Proteoglycan core protein expression, but not GAG biosynthetic enzyme expression, reflects embryonic lineage identity and EMT transcription factor expression in adult tissues

Given the central role of EMT in embryonic germ-layer formation and the established involvement of proteoglycans in this process (17, 30), we first investigated whether proteoglycan core protein and GAG biosynthetic enzyme expression in adult tissues are associated with embryonic lineage identity and EMT transcription factor (TF) expression. Single-cell RNA-sequencing (scRNA-seq) data from the Human Protein Atlas comprising 81 adult human cell types were analyzed to examine the expression of 42 proteoglycan core proteins and 48 GAG biosynthetic enzymes (**Additional file 2: Table S1A-B and S2**). Hierarchical clustering identified three cell type clusters (1–3) and three proteoglycan core protein clusters (A-C) (Fig. 1A). Examination of the resulting cell type clusters revealed a striking segregation according to embryonic germ-layer origin, with endoderm-derived cells predominantly clustering in cluster 1, mesoderm-derived cells in cluster 2, and ectoderm-derived cells in cluster 3 (Fig. 1A). This association was confirmed by Fisher's exact test (Fig. 1B), supporting the notion that embryonic germ-layer identity is retained in adult proteoglycan core protein expression patterns. The stability of the clusters and their cell types and gene compositions are shown in **Additional file 1: Figure S1A-F** and **Additional file 2: Table S2 and S3**.

Because EMT establishes embryonic germ-layer identity, we next asked whether the proteoglycan-defined clusters also differed in EMT TF expression. To address this, we compared the average expression of the major EMT TFs (*SNAI1/2*, *TWIST1/2*, and *ZEB1/2*) across the three clusters (Fig. 1A and 1C). The EMT TFs exhibited a similar clustering pattern. Cluster 2, which was enriched for mesodermal cell types, showed the highest overall expression of EMT TFs, with particularly high expression for *ZEB1/2* and *SNAI1*; cluster 3, corresponding to ectodermal cells, showed high expression of *ZEB1* and *ZEB2*, but not for any of the other EMT TFs; while cluster 1, dominated by endodermal cell types, exhibited generally lower expression of all EMT TFs, with the exception of *SNAI2*.

Despite the clear segregation by embryonic germ-layer origin, proteoglycan core proteins showed no obvious clustering according to their predominant GAG chain type, as indicated by the side-bar annotation (Fig. 1A). We therefore next investigated whether the enzymes responsible for GAG biosynthesis exhibited a similar lineage-associated organization. GAG biosynthesis is mediated by distinct enzyme classes responsible for chain initiation, polymerization, and modification through epimerization and sulfation. Unlike proteoglycan core proteins, GAG biosynthetic enzymes did not

segregate according to embryonic germ-layer origin (Fig. 1D). Instead, hierarchical clustering identified four cell-type clusters (1–4), with cluster stability and cell type assignments shown in **Additional file 1: Figure S1G-L** and **Additional file 2: Table S4 and S5**. Cluster 2 consisted exclusively of mesoderm-derived cell types and was highly enriched for immune cells, containing 11 of the 12 immune cell types included in the dataset. This cluster was primarily defined by enzymes belonging to GAG biosynthetic enzyme cluster C, many of which participate in chondroitin/dermatan sulfate biosynthesis (Fig. 1D). This enrichment suggests that immune cells may possess an increased capacity for chondroitin/dermatan sulfate biosynthesis. Cluster 4 was characterized by broadly elevated expression of GAG biosynthetic enzymes and consisted predominantly of principal parenchymal cells of the central nervous system (six of the eleven CNS cell types), suggesting a particular requirement for GAG-modified proteoglycans in the CNS. Cluster 1 consisted of three spermatogenic germ cell types and displayed generally lower expression of GAG biosynthetic enzymes (Fig. 1D). No distinct cluster corresponding to GAG chain initiation (linker) enzymes was identified. Instead, these enzymes were broadly expressed across cell types, in contrast to the remaining GAG biosynthetic enzyme classes, consistent with the widespread distribution of GAGs across human tissues (**Figure S1M**). Overall, enzymes involved in heparan, chondroitin, and dermatan sulfate biosynthesis exhibited marked cell type-specific expression but showed little association with embryonic germ-layer origin. We next examined whether the GAG biosynthetic enzyme clusters differed in EMT TF expression (Fig. 1D-E). Here, *ZEB1* and *ZEB2* were found to be highly expressed in the CNS-enriched cluster (cluster 4), whereas *SNAI1* and *ZEB2* expression was highest in the immune-enriched cluster (cluster 2). These findings suggest a weaker association between EMT TF activity and GAG biosynthetic enzyme expression than observed for proteoglycan core proteins.

In summary, proteoglycan core protein expression in adult tissues is strongly associated with embryonic germ-layer identity and EMT TF expression, whereas GAG biosynthetic enzymes exhibit more cell type- and tissue-specific expression patterns with only limited overlap with EMT TF expression. The observation that embryonic lineage remained the dominant organizing feature despite the diverse physiological functions of the adult cell types suggests that developmental expression programs may persist long after embryogenesis. These observations raised the possibility that the developmental transcriptional programs observed in adult tissues originate during embryogenesis. We therefore next examined proteoglycan core protein and GAG biosynthetic enzyme expression during human gastrulation.

Proteoglycan core proteins undergo dynamic regulation during gastrulation and partially retain embryonic germ-layer identity

To investigate whether the germ-layer-specific core protein patterns observed in adult tissues originate during embryogenesis, we analyzed scRNA-seq data from a karyotypically normal human embryo (gestational week 4 + 5 days) (31). In the analysis, we included only cells annotated as epiblast, primitive

streak, or the three germ layers, ectoderm, endoderm, mesoderm, as this is the stage of embryogenesis where EMT is first induced (Fig. 2A). To reconstruct the transcriptional changes accompanying germ-layer formation, we applied pseudotime trajectory analysis, which orders individual cells according to transcriptional similarity and identifies genes whose expression changes during lineage progression. The analysis reconstructed two developmental trajectories emerging from the epiblast, one leading to the endoderm and the other to the mesoderm (Fig. 2B). Association tests performed separately along each pseudotime trajectory identified multiple proteoglycan core proteins, GAG biosynthetic enzymes, and EMT TFs whose expression changed significantly during lineage progression (FDR < 0.05; **Additional file 3: Table S6A-B**). These associations indicate that the expression of these genes changes dynamically as cells transition through successive developmental states during germ-layer formation. Along the endoderm trajectory, 16 proteoglycan core proteins (38%), 12 GAG biosynthetic enzymes (25%), and three EMT TFs (50%) were significantly associated with pseudotime. Similar results were observed for the mesoderm trajectory, where 18 core proteins (43%), 13 GAG biosynthetic enzymes (27%), and five EMT TFs (83%) exhibited significant dynamic expression. To further characterize these dynamic expression changes, we examined the smoothed expression profiles of all significantly associated genes across the pseudotimes (**Additional file 1: Figure S2A-F**). Based on their smoothed expression profiles, genes were classified as displaying increasing, decreasing, or fluctuating expression patterns. For the endodermal trajectory, the majority of the significantly associated core proteins were either increasing (44%) or fluctuating (38%). In comparison, in the mesoderm trajectory, 56% of the significantly associated core proteins were increasing, and only 17% were fluctuating. As for the GAG biosynthetic enzymes, in the endodermal trajectory, the majority of the significantly associated GAG biosynthetic enzymes were increasing (67%), while for the mesoderm trajectory, the majority were fluctuating (62%). Interestingly, all the significantly associated EMT TFs had a fluctuating expression in the endodermal trajectory (*TWIST1*, *ZEB1*, and *ZEB2*) and either a fluctuating (60%; *SNAI1*, *TWIST1*, and *ZEB1*) or increasing (40%; *SNAI2* and *ZEB2*) expression in the mesodermal trajectory (**Additional file 1: Figure S2E and F**). The fluctuating expression of EMT TFs along the endodermal trajectory likely reflects the transient activation of EMT-associated transcriptional programs during definitive endoderm specification. Together, these findings indicate that proteoglycan core proteins, GAG biosynthetic enzymes, and EMT TFs undergo dynamic transcriptional changes during gastrulation, with distinct expression patterns observed along the endodermal and mesodermal trajectories.

Although both proteoglycan core proteins and GAG biosynthetic enzymes displayed dynamic expression during lineage progression, their individual expression patterns varied considerably. We therefore next asked whether these genes nevertheless exhibited coordinated expression with the EMT TFs. To address this, we performed a coordinated dynamics analysis. Rather than evaluating individual genes, this approach summarizes the average expression trajectory of predefined gene groups, thereby highlighting coordinated expression changes during lineage progression. Correlation analysis of these module trajectories revealed a negative EMT TF-core protein and EMT TF-GAG biosynthetic enzyme coupling in the endoderm trajectory ($r = -0.61$ and -0.61 , respectively), but a positive coupling in the mesoderm ($r = 0.41$ and 0.62), revealing lineage-specific relationships between EMT TF expression and both

proteoglycan core protein and GAG biosynthetic enzyme expression (Fig. 2C). Although the correlation was stronger for the GAG biosynthetic enzymes than the core proteins, fewer GAG biosynthetic enzymes were found to be significantly regulated during the pseudotime compared to the core proteins, as described in the trajectory analysis. Together, these data suggest that core protein and GAG biosynthetic enzyme dynamics are tightly coordinated with the EMT transcriptional program, but in lineage-specific and opposing directions during early human development.

To compare germ-layer specificity between embryonic development and adulthood while minimizing the impact of sparse single-cell expression, we first aggregated transcriptionally similar cells into metacells. For each gene, germ-layer specificity was defined as the germ layer with the highest expression among the top 75th percentile (Q3) of metacells in both embryonic and adult datasets (**Additional file 1: Figure S2G and H, Additional file 3: Table S8A-B**). For the adult tissue data, we did not use the heatmap cluster grouping but restricted the analysis to 74 cell types that could be confidently associated based on the literature with one of the three germ layers (**Additional file 3: Table S9**). In the embryonic scRNA-seq data, five proteoglycan core proteins were not detected and eighteen showed sparse expression with only a few metacells expressing them, thus nineteen proteoglycan core proteins were included in this analysis (Fig. 2D). Of these, nine (*AGRN*, *BGN*, *DCN*, *GPC5*, *LUM*, *SDC2*, *SDC4*, *TGFBR3*, and *TSKU*) maintained the same germ-layer specificity after cell type differentiation. In addition, *SDC3* exhibited equal Q3 expression in ectoderm and endoderm in embryonic development but shifted to an ectoderm-specific expression in adulthood. Interestingly, with the exception of *SDC4*, all of the proteoglycan core proteins showed a significant association with the lineage trajectories, for example *AGRN* with endoderm and *DCN*, *GPC5*, and *TGFBR3* with mesoderm (**Additional file 3: Table S6**). In contrast, nine proteoglycan core proteins changed their top-layer assignment: Four shifted from an embryonic epithelial layer (ectoderm or endoderm) to mesoderm in adulthood, one from endoderm to ectoderm, and four from mesoderm to either endoderm or ectoderm (Fig. 2D). Overall, 53% of proteoglycan core proteins retained their embryonic germ-layer specificity into adulthood. The combination of conserved germ-layer specificity and significant lineage associations during pseudotime suggests that embryonic developmental programs continue to shape proteoglycan core protein expression in adult tissues.

For the 48 GAG biosynthetic enzymes, ten showed sparse expression and two were not detected in the embryonic scRNA-Seq data. Among the remaining 36 enzymes, only eleven (*B4GALT7*, *CHST8*, *CHST9*, *CHST11*, *CSGALNACT2*, *DSEL*, *EXTL1*, *GLCE*, *HS6ST1*, *NDST1*, and *XYLT2*) retained their embryonic germ-layer specificity (31%), with four of these (*B4GALT7*, *CHST9*, *CHST11*, and *DSEL*) having a significant lineage association. By contrast, the other 25 enzymes shifted their germ-layer assignment: Nine from ectoderm or endoderm to mesoderm, five from mesoderm to ectoderm or endoderm, eight between ectoderm and endoderm, and one from dual embryonic (endoderm/mesoderm) to mesoderm in adulthood (Fig. 2E). Overall, only 31% of GAG biosynthetic enzymes retained their embryonic germ-layer specificity, substantially lower than the 53% observed for proteoglycan core proteins.

EMT transcription factors are associated with proteoglycan regulation and potential enhancer binding

The strong conservation of embryonic lineage specificity together with the coordinated expression of proteoglycan core proteins and EMT TFs during gastrulation suggested that EMT TFs may directly regulate the expression of specific proteoglycan core proteins. We therefore next examined the relationship between individual EMT TFs and proteoglycan core proteins, as well as GAG biosynthetic enzymes, using multiple complementary approaches. We first examined the relationship between individual EMT TFs and proteoglycan core proteins or GAG biosynthetic enzymes by calculating pairwise correlations between their smoothed pseudotime expression profiles within each lineage. Using a Pearson correlation threshold of $|r| \geq 0.4$, extensive correlation networks were identified in both the endodermal and mesodermal trajectories (Fig. 3A, **Additional file 3: Table S10A-B**). For the endodermal trajectory, 108 proteoglycan core protein-TF correlations were identified (46 positive, 62 negative), whereas the mesodermal trajectory yielded 140 correlations (71 positive, 69 negative). Interestingly, genes correlated with one EMT TF in the mesoderm trajectory almost invariably showed similar correlations with the remaining EMT TFs, suggesting a highly coordinated transcriptional program in this lineage. Further, the endoderm lineage displayed a modest enrichment for negative associations, implying distinct regulatory dynamics between the lineages. By comparison, the GAG biosynthetic enzyme analysis revealed an even broader regulatory network: 160 significant correlations in the endoderm (53 positive, 107 negative; **Fig. 3B, Additional file 3: Table S10C-D**) and 205 in the mesoderm (128 positive, 77 negative). However, the endoderm was skewed toward negative regulation, while the mesoderm was characterized by a predominance of positive TF-gene correlations. Strikingly, *SNAI1* stood out in the endoderm by showing mostly positive correlations with GAG biosynthetic enzymes, whereas the other EMT TFs maintained predominantly negative correlations. A similar yet more subtle *SNAI1*-specific trend was also evident in the proteoglycan core protein dataset, underscoring both shared and TF-specific regulatory programs across embryonic lineages. To further examine the correlation between EMT TF and proteoglycan core protein expression, we performed correlation analysis in the adult tissue data, using the same cutoff of $|r| \geq 0.4$ for defining a significant correlation (**Additional file 1: Figure S3A**). For the proteoglycan core proteins, 70 positive and 5 negative correlations were identified, while for the GAG biosynthetic enzymes 35 positive and 3 negative correlations were identified, mainly with the *ZEB1/2* TFs (Fig. 3C and D, **Additional file 3: Table S11A-B**). By examining the overlap between the adult tissue and embryonic data, we identified nineteen proteoglycan core proteins and eleven GAG biosynthetic enzymes (Fig. 3E and F). Although these correlations suggest coordinated regulation between EMT TFs and proteoglycan core proteins or GAG biosynthetic enzymes, they do not establish whether EMT TFs functionally regulate their expression. We therefore next analyzed publicly available datasets in which EMT TFs were perturbed by knockout (KO), knockdown (KD), or overexpression (OE).

A total of 44 publicly available *in vitro* EMT TF perturbation experiments were included in the analysis (**Additional file 4: Table S12**). Differential expression analysis was performed for each dataset and log2

fold changes (LogFC) were calculated between perturbed and wild-type samples (**Additional file 4: Table S12**). Because several datasets contained only a single biological replicate, we focused on the direction and magnitude of expression changes rather than statistical significance. To enable consistent interpretation, LogFC values from KO and KD experiments were inverted so that all changes could be interpreted in the same direction as OE, thereby reflecting the direct effect of the TF, where a positive value indicates positive regulation by the EMT TFs, and a negative value reflects a negative regulation. To account for differences in experimental design between datasets, we summarized the merged LogFC distributions using the upper quartile (Q3), thereby capturing the predominant direction of regulation across studies. We applied a LogFC threshold of ± 1 (corresponding to a two-fold change) as a heuristic to identify biologically meaningful shifts. EMT TF perturbation altered the expression of 23 proteoglycan core proteins (~ 55%), with ZEB family TFs accounting for approximately 70% of the observed regulatory events (Fig. 3G, **Additional file 1: Figure S3B, Additional file 4: Table S12**). Among the GAG biosynthetic enzymes, 14 (~ 29%) exhibited altered expression following EMT TF perturbation, with an even greater predominance of ZEB family TFs (~ 86%) (Fig. 3H, **Additional file 1: Figure S3B, Additional file 4: Table S12**). Among the genes meeting the predefined threshold, most regulatory effects were positive for both proteoglycan core proteins (68%) and GAG biosynthetic enzymes (63%).

Although perturbation experiments provide evidence that EMT TFs influence gene expression, they cannot distinguish between direct and indirect regulatory effects. We therefore investigated whether EMT TFs bind regulatory regions associated with proteoglycan core protein and GAG biosynthetic enzyme genes using publicly available ChIP-seq datasets. The analysis was restricted to the ZEB and SNAIL families because limited ChIP-seq data were available for the TWIST family (**Additional file 5: Table S13**). Information on enhancer and promoter regions of proteoglycan core protein and GAG biosynthetic genes were obtained from GeneHancer (<https://genehancer-staging.azurewebsites.net>) (**Additional file 5: Table S14**). To identify robust interactions across multiple biological contexts, an EMT TF was considered a potential regulator only when binding to the same GeneHancer element was detected in at least 50% of the available datasets. Using this criterion, we found that fifteen (36%) proteoglycan core proteins exhibited binding of at least one EMT TF through the same regulatory element (Fig. 3I, **Additional file 5: Table S15A-B**). Binding was most frequent for ZEB1 and SNAIL1, potentially regulating seven and six proteoglycan core proteins, respectively, followed by ZEB2 (four) and SNAIL2 (three). Notably, all interactions were located within enhancer elements, with none detected in promoter regions. Among the GAG biosynthetic enzymes, twenty-four (50%) showed potential binding of at least one EMT TF in their promoter or enhancer regions. Interestingly, a predominance of ZEB-family TFs was observed, with both ZEB1 and ZEB2 binding the regulatory regions of nineteen GAG enzymes, followed by SNAIL1 (eight) and SNAIL2 (five) (Fig. 3J, **Additional file 5: Table S15C-D**). In contrast to the proteoglycan core proteins, four interactions were identified within promoter elements, two of each by ZEB1 and SNAIL1, while the majority were still associated with enhancers. Together with the perturbation analyses, these findings support direct regulation of numerous proteoglycan core proteins and GAG biosynthetic enzymes by EMT TFs, predominantly through enhancer-associated binding sites.

To integrate the findings across all analyses, we compiled a gene-level summary of adult tissue expression, embryonic trajectory associations, EMT TF correlations, perturbation experiments, and ChIP-seq evidence for every proteoglycan core protein and GAG biosynthetic enzyme (**Additional file 6: Table S16**).

Discussion

Our study demonstrates that proteoglycan core protein expression in adult tissues retains a strong association with embryonic developmental origin despite extensive tissue differentiation. By integrating adult and embryonic single-cell transcriptomic datasets with EMT TF perturbation and ChIP-seq analyses, we show that proteoglycan core proteins retain substantially stronger conservation of embryonic germ-layer specificity than GAG biosynthetic enzymes and identify EMT-associated transcriptional programs as candidate regulators of this developmental memory. Together, these findings suggest that the adult proteoglycan expression landscape is shaped, at least in part, by transcriptional programs established during gastrulation, thereby extending the current understanding of proteoglycan biology beyond their well-established structural and signaling functions.

Gastrulation is the first developmental process during which EMT establishes the three embryonic germ layers and therefore represents a critical stage for investigating how developmental transcriptional programs are initiated. Proteoglycans are already known to play essential roles during this process by regulating signaling pathways required for lineage specification and tissue morphogenesis. For example, proteoglycans are required for fibroblast growth factor (FGF) signaling during primitive streak formation in mouse embryos (32), while deletion of the heparan sulfate polymerizing enzymes EXT1 or EXT2 results in failure of gastrulation (33). Another well-characterized EMT-associated developmental process is the formation of the neural crest, a transient migratory cell population that expresses EMT TFs including SNAI1, SNAI2, and TWIST1 (34–37). Although neural crest cells originate from the ectoderm, their migration has been proposed to be guided by proteoglycan expression within the sclerotome, a mesoderm-derived tissue (38). In support of this, a study in rat embryos demonstrated that enzymatic digestion of chondroitin and dermatan sulfate chains using chondroitinase ABC inhibited neural crest cell migration (39). The authors proposed that chondroitin/dermatan sulfate proteoglycans facilitate migration by reducing extracellular matrix adhesiveness, thereby creating a permissive environment for migrating neural crest cells. Similar observations have subsequently been reported in mouse embryos (40).

Although the developmental functions of proteoglycans are well established, considerably less is known about how their expression is regulated during embryogenesis. Our analyses demonstrate that both proteoglycan core proteins and GAG biosynthetic enzymes undergo dynamic transcriptional changes throughout endodermal and mesodermal lineage specification that closely parallel EMT TF expression. GAG biosynthetic enzymes displayed predominantly positive correlations with EMT TFs along the mesoderm trajectory but predominantly negative correlations along the endoderm trajectory, suggesting distinct lineage-specific regulatory programs during germ-layer formation.

An additional observation supporting lineage-specific regulation was the markedly different temporal expression dynamics observed along the two developmental trajectories. Whereas mesoderm-associated proteoglycan core proteins and EMT TFs predominantly displayed increasing expression throughout pseudotime, a substantially larger proportion of proteoglycan core proteins, GAG biosynthetic enzymes, and all significantly associated EMT TFs exhibited fluctuating expression along the endoderm trajectory. Rather than reflecting stable acquisition of a mesenchymal phenotype, these dynamic expression patterns are consistent with the transient epithelial plasticity that accompanies endoderm specification during gastrulation (41). The fluctuating expression of multiple proteoglycan core proteins and GAG biosynthetic enzymes therefore suggests that remodeling of the proteoglycan landscape accompanies these transient changes in cellular state, indicating that proteoglycans may contribute not only to establishing stable epithelial and mesenchymal phenotypes but also to the epithelial plasticity that accompanies developmental cell-state transitions.

Despite this coordinated regulation during embryogenesis, these relationships were not uniformly maintained after development. Whereas coordinated expression between EMT TFs and proteoglycan core proteins persisted into adult tissues, the widespread embryonic associations observed for GAG biosynthetic enzymes were largely lost after development, with the exception of mesoderm-derived immune cells and the ZEB family of EMT TFs. This difference likely reflects the fundamentally distinct biological contexts of gastrulation and adult tissue homeostasis. One possibility is that proteoglycan core proteins and GAG biosynthetic enzymes are co-regulated by additional developmental pathways that are activated concurrently with EMT during embryogenesis, resulting in broad transcriptional coordination. Alternatively, EMT-dependent regulation may be transient, functioning primarily during lineage specification before tissue-specific regulatory mechanisms become dominant in differentiated adult cells. These mechanisms are not mutually exclusive and together may explain why only a subset of embryonic EMT-associated correlations persists into adulthood. Finally, although our embryonic analyses were limited to a single human gastrulation dataset and the inherent sparsity of single-cell RNA sequencing likely reduced detection of some genes, the consistency between the embryonic trajectory analyses, adult tissue datasets, EMT TF perturbation experiments, and ChIP-seq analyses supports the robustness of the developmental relationships identified here.

One of the clearest patterns identified in the present study was the segregation of proteoglycan core protein expression according to embryonic germ-layer origin. This observation is consistent with the well-established developmental relationship between germ-layer origin and the predominant characteristics of differentiated cell types. For example, the endoderm gives rise primarily to epithelial linings, including those of the gastrointestinal tract, respiratory system, and urinary bladder, whereas the mesoderm gives rise to connective tissue and stromal cell types, including fibroblasts, smooth muscle cells, and many immune cells, that frequently display mesenchymal characteristics. Consistent with these developmental relationships, EMT TF expression was lower in the endoderm-enriched cluster and higher in the mesoderm-enriched cluster. Because the clustering was based solely on proteoglycan core protein expression, these findings suggest that subsets of proteoglycans are preferentially associated with cell types of endodermal or mesodermal origin. Given the predominant epithelial and mesenchymal

characteristics of these germ layers, respectively, this may reflect differences in the proteoglycan repertoires associated with these broad cellular phenotypes.

Among the proteoglycan core proteins most strongly associated with mesoderm-derived cell types and mesenchymal characteristics were the three small leucine-rich proteoglycans biglycan (BGN), decorin (DCN), and lumican (LUM). These secreted extracellular matrix proteoglycans are key regulators of collagen fibrillogenesis and extracellular matrix organization, consistent with their high expression in stromal cell populations in our dataset, including fibroblasts, endometrial stromal cells, ovarian stromal cells, and Leydig cells. Their association with mesoderm-derived tissues is further supported by the well-established roles of biglycan and decorin in skeletal development and bone homeostasis (42), with biglycan in particular being required for normal bone formation during embryonic development (43–45). The observed expression pattern is also consistent with previous studies linking all three proteoglycans to EMT. Biglycan, decorin, and lumican have all been included as mesenchymal markers in published EMT signatures (17), while decorin and lumican have been shown to be upregulated during TGF- β -induced EMT (46). Furthermore, biglycan has repeatedly been associated with EMT and metastatic progression in several cancer types (47–51). Interestingly, despite their association with mesenchymal phenotypes, both lumican and decorin have also been reported to suppress EMT and cell migration in cancer (52–55), highlighting the context-dependent functions of proteoglycans in EMT regulation.

Another proteoglycan showing a strong association with mesoderm-derived cell types was syndecan-2 (SDC2). Consistent with our findings, syndecan-2 has previously been reported to be selectively expressed in mesenchymal precursor cells during mouse embryonic development (56). In addition, syndecan-2 is upregulated during TGF- β -induced EMT (46), has been identified as a mesenchymal marker in published EMT signatures (17), and promotes EMT and migration in colorectal cancer cells (57, 58). Although the association of individual proteoglycan core proteins with epithelial or mesenchymal characteristics varied across our analyses, the overall pattern was consistent with preferential expression of several extracellular matrix-associated proteoglycans in mesoderm-derived tissues, many of which also retained their embryonic germ-layer specificity into adulthood.

In contrast to the proteoglycan core proteins associated with mesoderm-derived cell types, syndecan-1 (SDC1) and syndecan-4 (SDC4) were among the most highly expressed proteoglycans in endoderm-derived cell types. Consistent with this observation, both have previously been included in epithelial-associated EMT signatures (17). Syndecan-1 is predominantly expressed by epithelial cells, where it contributes to anchoring cells to the extracellular matrix (59), and syndecan-4 has likewise been associated with an epithelial phenotype following TGF- β -induced EMT (46). However, the role of syndecan-4 appears to be context-dependent, as it has also been reported to promote EMT and cell migration in other experimental systems (60, 61). This complexity was also reflected in our analyses. Although *SDC1* and *SDC4* were enriched in endoderm-derived cell types, neither exhibited strict endoderm-specific expression across the datasets analyzed, nor did we find evidence for consistent negative regulation by EMT TFs. Together, these findings suggest that, unlike several mesenchymal-

associated proteoglycan core proteins, the relationship between syndecans and EMT is likely to depend on cellular context rather than a simple epithelial–mesenchymal dichotomy.

Our integrative analyses suggest that EMT TFs are candidate regulators of a substantial proportion of proteoglycan core proteins and, to a lesser extent, GAG biosynthetic enzymes. By combining embryonic and adult correlation analyses with perturbation and ChIP-seq datasets, we identified multiple proteoglycan core proteins that consistently showed evidence of association with EMT TF activity. Although these analyses do not establish direct regulation for every candidate, they provide a systematic framework for prioritizing proteoglycan core proteins and GAG biosynthetic enzymes for future experimental validation. Previous studies investigating direct regulation of proteoglycan expression by EMT TFs have largely focused on individual proteoglycans. For example, ZEB1 has been shown to bind the syndecan-1 promoter and repress its expression in prostate cancer cells (62), while putative SNAI1 binding sites have been proposed in the promoters of syndecan-1 and syndecan-2, although without direct experimental validation (63). Consistent with these observations, we identified a negative correlation between *ZEB1* and *SDC1* expression in adult tissues, but found no additional evidence supporting direct regulation across our complementary analyses. Likewise, TWIST1 has been reported to repress the *Aggrn* promoter during mouse chondrogenesis (64), while overexpression of *Snai1* and *Snai2* also reduces *Aggrn* expression, although direct binding was not investigated (65). In contrast, our analyses classified *AGRN* as an endoderm-associated proteoglycan core protein but did not support regulation by EMT TFs. Finally, SNAI1 has been shown to directly activate *VCAN* expression in breast cancer cells (66). Although we also identified associations between *VCAN* and EMT TFs, both positive and negative relationships were observed across our datasets, suggesting that *VCAN* regulation is likely influenced by additional context-dependent regulatory mechanisms.

Interestingly, several studies have demonstrated reciprocal regulation, whereby proteoglycans themselves influence EMT TF expression. For example, knockdown of syndecan-2 reduced SNAI2 and TWIST protein expression in colorectal cancer cells (57), loss of biglycan decreased expression of ZEB1, ZEB2, and SNAI1 in gastric cancer cells (49), and silencing of SPOCK1 reduced expression of SNAI1, SNAI2, ZEB1, and ZEB2 in pancreatic cancer cells (67). Collectively, these observations suggest that regulatory interactions between proteoglycans and EMT TFs may be bidirectional, with proteoglycans functioning not only as downstream targets of EMT-associated transcriptional programs but also as upstream modulators of EMT TF activity.

Several limitations should nevertheless be considered when interpreting the EMT TF perturbation and regulatory binding analyses. The publicly available perturbation datasets were limited in number and predominantly derived from cancer cell lines, introducing both biological and technical heterogeneity. Moreover, EMT is increasingly recognized as a highly context-dependent process that varies across developmental stages, tissues, and disease settings (68–70). Consequently, regulation of proteoglycan core proteins and GAG biosynthetic enzymes by EMT TFs is likely to be similarly context-dependent. Although integrating multiple independent datasets substantially strengthened the robustness of our

conclusions, experimental validation in defined developmental and physiological models will be required to establish direct regulatory relationships.

Aberrant activation of EMT contributes to pathological tissue remodeling in numerous diseases, including chronic fibrotic disorders such as pulmonary fibrosis and liver cirrhosis, as well as tumor invasion, metastasis, and therapy resistance in cancer. Although the present study focused on developmental EMT rather than disease-specific contexts, understanding how proteoglycan core proteins and GAG biosynthetic enzymes are regulated during embryonic development provides an important framework for interpreting their dysregulation when EMT is aberrantly activated. The observation that multiple proteoglycan core proteins and GAG biosynthetic enzymes exhibit dynamic expression during epithelial plasticity further raises the possibility that these molecules may also contribute to the partial EMT states that characterize cancer progression and therapy resistance. Our findings suggest that proteoglycan core proteins retain aspects of their developmental regulation into adulthood and identify candidate EMT TF–proteoglycan regulatory relationships that may become reactivated during pathological EMT. These observations provide a rational basis for prioritizing proteoglycan core proteins and GAG biosynthetic enzymes for future functional studies in fibrosis, cancer, and other diseases involving epithelial and mesenchymal plasticity. More broadly, this study illustrates the value of integrating multiple publicly available transcriptomic, perturbation, and TF-binding datasets to identify conserved developmental regulatory programs that would be difficult to resolve using individual datasets alone.

Conclusions

This study identifies a previously unrecognized developmental relationship between EMT transcriptional programs and proteoglycan core protein expression. By integrating adult and embryonic single-cell transcriptomic data with EMT TF perturbation and ChIP-seq analyses, we demonstrate that proteoglycan core proteins retain stronger signatures of embryonic germ-layer identity and coordinated EMT TF activity than GAG biosynthetic enzymes, which instead exhibit predominantly cell type- and tissue-specific expression patterns. Together, these findings suggest that the adult proteoglycan landscape is shaped, at least in part, by transcriptional programs established during embryogenesis and identify EMT TFs as candidate regulators of this developmental memory. The integrative framework presented here provides a resource for prioritizing proteoglycan core proteins and GAG biosynthetic enzymes for future mechanistic studies of development, tissue homeostasis, regeneration, fibrosis, and cancer.

Methods

Data from Human Protein Atlas

Single-cell RNA expression data of normal adult tissue were obtained from the Human Protein Atlas (HPA) database (<https://www.proteinatlas.org>) under the downloadable data section for single cell types. The dataset comprised transcript-per-million (TPM)-normalized RNA expression values for 81

human cell types aggregated from 31 studies. Germ-layer origins were assigned based on published literature and independently verified using AI-assisted literature cross-referencing. Expression values were $\log_2$ -transformed prior to downstream analyses.

Proteoglycan core protein and glycosaminoglycan (GAG) biosynthetic enzymes gene lists (**Additional file 2: Table S1A-B**) were curated, and GAG modification patterns for individual proteoglycans were annotated based on previously published studies (71, 72).

Hierarchical clustering

Hierarchical clustering was performed on $\log_2$ -transformed expression values using the ComplexHeatmap v2.20.0 package in R (73). Cluster stability was evaluated independently for proteoglycan core proteins and GAG biosynthetic enzymes using clValid v0.7, assessing connectedness, Dunn index, and silhouette width across 2–10 cluster solutions to determine the optimal number of clusters (74). Visualization was performed with ComplexHeatmap, RColorBrewer, circlize, ggpubr, and ggthemes. The distribution of germ-layer assignments across clusters was tested using Fisher's exact test.

Correlation analysis

Pearson correlation coefficients between EMT TFs and proteoglycan core protein or GAG biosynthetic enzyme expression were calculated from $\log_2$ -transformed expression values using the `cor()` function in the R stats package.

Single-cell RNA-seq data analysis

Processed single-cell RNA-sequencing (scRNA-seq) data of a karyotypically normal male human embryo at gestational week 4 + 5 days (Carnegie Stage 7; CS7, equivalent to 16–19 days post-fertilization) were obtained from a publicly available dataset (31). Only cells annotated as Epiblast, Primitive Streak, Endoderm, Mesoderm, or Ectoderm were retained, corresponding to the developmental window when EMT is first initiated. Cluster identities were remapped to a Subset_Cell_Types factor, and the dataset was reprocessed in Seurat v5 (75). Normalization, variable-feature selection, and scaling were performed using SCTransform with default parameters. Dimensionality reduction was carried out with PCA, and visualization was done with UMAP (dims 1–20). Cell-type annotations were adopted from the original publication.

Metacell analysis

Metacells were generated using the hdWGCNA package with the fraction method, including genes expressed in at least 5% of cells (76). Cells were grouped by cell types, with parameters $k = 10$,

max_shared = 5, and min_cells = 20. Metacells were normalized with SCTransform and log-normalized expression of genes in metacells was visualized using SCpubr (77).

Pseudotime trajectory analysis

Developmental trajectories were reconstructed using Monocle3 in conjunction with Seurat objects. The Epiblast was defined as the root by identifying the closest principal-graph vertex, and cells were ordered along pseudotime with order_cells(). Lineage branches were isolated using choose_graph_segments() to define endodermal and mesodermal trajectories.

Trajectory and gene-association analysis

Raw UMI counts from the Seurat RNA assay were modeled as a smooth function of pseudotime using tradeSeq v1.16 (*fitGAM*, 6 knots) (78). Genes with Benjamini-Hochberg-adjusted $p < 0.05$ in *associationTest* were considered significant. Analyses included 42 proteoglycan core proteins, 39 GAG biosynthetic enzymes, and six EMT-related TFs (SNAI1/2, TWIST1/2, ZEB1/2).

Module correlation

Smoothed expression curves were generated with *predictSmooth* ($n = 200$ points), from tradeSeq, and z-scored per gene. Mean module trajectories were calculated for proteoglycan core proteins, GAG biosynthetic enzymes, and EMT TFs, and Pearson correlations were computed between modules within each lineage using custom R functions.

Bulk RNA-seq data analysis

Raw sequencing data were downloaded from publicly available bulk RNA-seq data on EMT TF perturbation experiments (knockout, knockdown, or overexpressed) (**Additional file 4: Table S12**). Raw fastq files were adapter trimmed, aligned to human reference genome GRCh38 v104 and raw count matrices were generated using CLC Genomics Workbench 22.

Each BioProject was analyzed separately to preserve dataset-specific batch structures. For projects containing multiple biological replicates per condition, differential expression analysis was conducted using the edgeR-limma workflow. Gene counts were first filtered using filterByExpr to remove genes with consistently low expression across all samples, followed by Trimmed Mean of M-values (TMM) normalization implemented in calcNormFactors (edgeR). Library sizes and normalization factors were used to generate log-transformed counts per million (logCPM) values. Quality assessment and sample relationships were visualized by multidimensional scaling (MDS) based on the logCPM matrix, and the samples showing evidence of batch effects or other confounding variation were added to the design matrix to address batch effect. Normalized counts were then analyzed with the voom-limma pipeline, using quality weights to account for between-sample variability. Linear models were fitted per gene, and empirical Bayes moderation of variance was applied to obtain the final differential expression analysis result. Contrasts were defined according to experimental conditions (e.g., overexpression vs wild-type, knockout vs control), and Benjamini-Hochberg correction was used to adjust p-values for multiple

testing. The resulting differential expression tables included $\log_2$ -fold change (logFC), average expression, t-statistics, and adjusted p-values for all detected genes.

For datasets lacking biological replicates or having single samples per condition, a simplified comparison was performed to estimate expression changes. Counts were first pseudocounted (+ 0.1) to avoid undefined log values, followed by direct computation of per-gene $\log_2$ -fold changes between experimental and control samples.

ChIP-seq data analysis

Raw ChIP-seq reads were aligned to the human reference genome GRCh38 (Ensembl v104) using CLC Genomics Workbench (Qiagen) with default parameters, retaining only uniquely mapped reads. The resulting BAM files were processed with samtools to retain reads with mapping quality ≥ 20 and, for paired-end libraries, only properly paired reads. Duplicate-retained and deduplicated versions were generated using samtools markdup to assess signal quality under both conditions. Peaks were called using MACS3 with a q-value threshold of 0.01 and the parameters `--nomodel --extsize 147` for single-end data. ENCODE hg38 v2 blacklist regions were removed with bedtools, and standard quality metrics, including the Fraction of Reads in Peaks (FRiP) score and total peak count, were computed. Samples were retained if FRiP ≥ 0.05 or if at least 2000 blacklist-filtered peaks were detected, ensuring adequate enrichment relative to background noise.

Downstream analyses were conducted using a custom R workflow to identify transcription-factor (TF) binding events associated with proteoglycan core protein and GAG biosynthetic enzymes genes. Briefly, narrowPeak files were imported as GRanges objects, preserving genomic coordinates and peak attributes. Annotated regulatory regions containing promoter and enhancer elements linked to proteoglycan core protein and GAG biosynthetic enzymes genes were obtained from the GeneHancer database (79) and harmonized to the same genome assembly. Genomic overlaps between TF peaks and regulatory elements were identified using the `findOverlaps()` function (GenomicRanges package), enabling the identification of candidate direct TF-target interactions.

Declarations

Ethics approval and consent to participate

This study used publicly available, previously published datasets. No new human participants, human tissue samples, or primary experimental data were generated as part of this study. Ethical approval and consent procedures for the original datasets were described in the corresponding original publications. Additional ethical approval was not required for the present secondary analysis.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflicts of interest with the contents of this article.

Funding

This study was funded by UiT The Arctic University of Norway.

Author Contribution

SOK, SVH, and EK conceived the study. SDS, AKMB, and EK designed the study and developed the methodology. SDA, AKMB, TNKQ, and BH performed bioinformatic and statistical analyses. SDA, AKMB, TNKQ, BH, SVH, and EK interpreted the data. CLRM and SOL provided domain-specific expertise in proteoglycan biology, developmental biology, and glycobiology. SDA, AKMB, and TNKQ prepared the figures and performed data visualization. SDS, KS, SVH, and EK wrote the original manuscript draft. All authors critically reviewed, edited, and approved the final version of the manuscript.

Acknowledgement

We sincerely thank the investigators and study participants who generated and made their datasets publicly available. This work would not have been possible without their commitment to open science and data sharing. We also acknowledge the public repositories and databases that curate and maintain these valuable resources.

Data Availability

No new sequencing or primary experimental datasets were generated in this study. Publicly available datasets analyzed in this study were obtained from the original repositories and resources described in the Methods section. Processed data files and derived summary tables generated during the analyses will be deposited in Zenodo [DOI to be added]. The R scripts used for data processing, analysis, and visualization are available on GitHub at https://github.com/UiT-Tumorbiology/Embryo_EMT_PG under the MIT License.

References

1. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419.
2. Iozzo RV, Schaefer L. Proteoglycan form and function: A comprehensive nomenclature of proteoglycans. *Matrix Biol*. 2015;42:11-55.

3. Denzer AJ, Brandenberger R, Gesemann M, Chiquet M, Ruegg MA. Agrin binds to the nerve-muscle basal lamina via laminin. *J Cell Biol.* 1997;137(3):671-83.
4. LeBaron RG, Zimmermann DR, Ruoslahti E. Hyaluronate binding properties of versican. *J Biol Chem.* 1992;267(14):10003-10.
5. Bengtsson E, Aspberg A, Heinegard D, Sommarin Y, Spillmann D. The amino-terminal part of PRELP binds to heparin and heparan sulfate. *J Biol Chem.* 2000;275(52):40695-702.
6. Bengtsson E, Mörgelin M, Sasaki T, Timpl R, Heinegård D, Aspberg A. The leucine-rich repeat protein PRELP binds perlecan and collagens and may function as a basement membrane anchor. *J Biol Chem.* 2002;277(17):15061-8.
7. Haglund L, Tillgren V, Onnerfjord P, Heinegard D. The C-terminal peptide of chondroadherin modulates cellular activity by selectively binding to heparan sulfate chains. *J Biol Chem.* 2013;288(2):995-1008.
8. Heinegård D. Fell-Muir Lecture: Proteoglycans and more--from molecules to biology. *Int J Exp Pathol.* 2009;90(6):575-86.
9. Stoler-Barak L, Moussion C, Shezen E, Hatzav M, Sixt M, Alon R. Blood vessels pattern heparan sulfate gradients between their apical and basolateral aspects. *PLoS One.* 2014;9(1):e85699.
10. Trelstad RL, Hayashi A, Hayashi K, Donahoe PK. The epithelial-mesenchymal interface of the male rat Mullerian duct: loss of basement membrane integrity and ductal regression. *Dev Biol.* 1982;92(1):27-40.
11. Santra M, Reed CC, Iozzo RV. Decorin binds to a narrow region of the epidermal growth factor (EGF) receptor, partially overlapping but distinct from the EGF-binding epitope. *J Biol Chem.* 2002;277(38):35671-81.
12. Berendsen AD, Pinnow EL, Maeda A, Brown AC, McCartney-Francis N, Kram V, et al. Biglycan modulates angiogenesis and bone formation during fracture healing. *Matrix Biol.* 2014;35:223-31.
13. Corti F, Wang Y, Rhodes JM, Atri D, Archer-Hartmann S, Zhang J, et al. N-terminal syndecan-2 domain selectively enhances 6-O heparan sulfate chains sulfation and promotes VEGFA(165)-dependent neovascularization. *Nat Commun.* 2019;10(1):1562.
14. Camper L, Heinegard D, Lundgren-Akerlund E. Integrin alpha2beta1 is a receptor for the cartilage matrix protein chondroadherin. *J Cell Biol.* 1997;138(5):1159-67.
15. Tillgren V, Onnerfjord P, Haglund L, Heinegård D. The tyrosine sulfate-rich domains of the LRR proteins fibromodulin and osteoadherin bind motifs of basic clusters in a variety of heparin-binding proteins, including bioactive factors. *J Biol Chem.* 2009;284(42):28543-53.
16. Kirkbride KC, Townsend TA, Bruinsma MW, Barnett JV, Blobe GC. Bone morphogenetic proteins signal through the transforming growth factor-beta type III receptor. *J Biol Chem.* 2008;283(12):7628-37.
17. Karagiorgou Z, Fountas PN, Manou D, Knutsen E, Theocharis AD. Proteoglycans Determine the Dynamic Landscape of EMT and Cancer Cell Stemness. *Cancers (Basel).* 2022;14(21).

18. Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol.* 2014;15(3):178-96.
19. Ridley AJ. Life at the leading edge. *Cell.* 2011;145(7):1012-22.
20. French MM, Smith SE, Akanbi K, Sanford T, Hecht J, Farach-Carson MC, et al. Expression of the heparan sulfate proteoglycan, perlecan, during mouse embryogenesis and perlecan chondrogenic activity in vitro. *J Cell Biol.* 1999;145(5):1103-15.
21. Yan D, Lin X. Shaping morphogen gradients by proteoglycans. *Cold Spring Harb Perspect Biol.* 2009;1(3):a002493.
22. Chen J, Sun T, You Y, Wu B, Wang X, Wu J. Proteoglycans and Glycosaminoglycans in Stem Cell Homeostasis and Bone Tissue Regeneration. *Front Cell Dev Biol.* 2021;9:760532.
23. Kirn-Safran CB, D'Souza SS, Carson DD. Heparan sulfate proteoglycans and their binding proteins in embryo implantation and placentation. *Semin Cell Dev Biol.* 2008;19(2):187-93.
24. García-García MJ, Anderson KV. Essential role of glycosaminoglycans in Fgf signaling during mouse gastrulation. *Cell.* 2003;114(6):727-37.
25. Brickman MC, Gerhart JC. Heparitinase inhibition of mesoderm induction and gastrulation in *Xenopus laevis* embryos. *Dev Biol.* 1994;164(2):484-501.
26. Ohkawara B, Yamamoto TS, Tada M, Ueno N. Role of glypican 4 in the regulation of convergent extension movements during gastrulation in *Xenopus laevis*. *Development.* 2003;130(10):2129-38.
27. Muñoz R, Moreno M, Oliva C, Orbenes C, Larraín J. Syndecan-4 regulates non-canonical Wnt signalling and is essential for convergent and extension movements in *Xenopus* embryos. *Nat Cell Biol.* 2006;8(5):492-500.
28. Zamir EA, Rongish BJ, Little CD. The ECM moves during primitive streak formation--computation of ECM versus cellular motion. *PLoS Biol.* 2008;6(10):e247.
29. Nakaya Y, Sukowati EW, Wu Y, Sheng G. RhoA and microtubule dynamics control cell-basement membrane interaction in EMT during gastrulation. *Nat Cell Biol.* 2008;10(7):765-75.
30. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. *J Clin Invest.* 2009;119(6):1420-8.
31. Tyser RCV, Mahammadov E, Nakanoh S, Vallier L, Scialdone A, Srinivas S. Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature.* 2021;600(7888):285-9.
32. Cao D, Bergmann J, Zhong L, Hemalatha A, Dingare C, Jensen T, et al. Selective utilization of glucose metabolism guides mammalian gastrulation. *Nature.* 2024;634(8035):919-28.
33. Lin X, Wei G, Shi Z, Dryer L, Esko JD, Wells DE, et al. Disruption of gastrulation and heparan sulfate biosynthesis in EXT1-deficient mice. *Dev Biol.* 2000;224(2):299-311.
34. Aybar MJ, Nieto MA, Mayor R. Snail precedes slug in the genetic cascade required for the specification and migration of the *Xenopus* neural crest. *Development.* 2003;130(3):483-94.
35. Nieto MA, Sargent MG, Wilkinson DG, Cooke J. Control of cell behavior during vertebrate development by Slug, a zinc finger gene. *Science.* 1994;264(5160):835-9.

36. Zhai J, Guo J, Wan H, Qi L, Liu L, Xiao Z, et al. Primate gastrulation and early organogenesis at single-cell resolution. *Nature*. 2022;612(7941):732-8.
37. Martik ML, Bronner ME. Riding the crest to get a head: neural crest evolution in vertebrates. *Nat Rev Neurosci*. 2021;22(10):616-26.
38. Dutt S, Kléber M, Matasci M, Sommer L, Zimmermann DR. Versican V0 and V1 guide migratory neural crest cells. *J Biol Chem*. 2006;281(17):12123-31.
39. Morriss-Kay G, Tuckett F. Immunohistochemical localisation of chondroitin sulphate proteoglycans and the effects of chondroitinase ABC in 9- to 11-day rat embryos. *Development*. 1989;106(4):787-98.
40. Kubota Y, Morita T, Kusakabe M, Sakakura T, Ito K. Spatial and temporal changes in chondroitin sulfate distribution in the sclerotome play an essential role in the formation of migration patterns of mouse neural crest cells. *Dev Dyn*. 1999;214(1):55-65.
41. Scheibner K, Schirge S, Burtscher I, Buttner M, Sterr M, Yang D, et al. Epithelial cell plasticity drives endoderm formation during gastrulation. *Nat Cell Biol*. 2021;23(7):692-703.
42. Fisher LW, Termine JD, Young MF. Deduced protein sequence of bone small proteoglycan I (biglycan) shows homology with proteoglycan II (decorin) and several nonconnective tissue proteins in a variety of species. *J Biol Chem*. 1989;264(8):4571-6.
43. Shainer R, Kram V, Kilts TM, Li L, Doyle AD, Shainer I, et al. Biglycan regulates bone development and regeneration. *Front Physiol*. 2023;14:1119368.
44. Xu T, Bianco P, Fisher LW, Longenecker G, Smith E, Goldstein S, et al. Targeted disruption of the biglycan gene leads to an osteoporosis-like phenotype in mice. *Nat Genet*. 1998;20(1):78-82.
45. Hua R, Han Y, Ni Q, Fajardo RJ, Iozzo RV, Ahmed R, et al. Pivotal roles of biglycan and decorin in regulating bone mass, water retention, and bone toughness. *Bone Res*. 2025;13(1):2.
46. Tellez-Gabriel M, Tekpli X, Reine TM, Hegge B, Nielsen SR, Chen M, et al. Serglycin Is Involved in TGF- β Induced Epithelial-Mesenchymal Transition and Is Highly Expressed by Immune Cells in Breast Cancer Tissue. *Front Oncol*. 2022;12:868868.
47. Schulz GB, Grimm T, Sers C, Riemer P, Elmasry M, Kirchner T, et al. Prognostic value and association with epithelial-mesenchymal transition and molecular subtypes of the proteoglycan biglycan in advanced bladder cancer. *Urol Oncol*. 2019;37(8):530.e9-.e18.
48. Yang Y, Wang R, Feng L, Ma H, Fang J. LINC00460 Promotes Cell Proliferation, Migration, Invasion, and Epithelial-Mesenchymal Transition of Head and Neck Squamous Cell Carcinoma via miR-320a/BGN Axis. *Onco Targets Ther*. 2021;14:2279-91.
49. Pinto F, Santos-Ferreira L, Pinto MT, Gomes C, Reis CA. The Extracellular Small Leucine-Rich Proteoglycan Biglycan Is a Key Player in Gastric Cancer Aggressiveness. *Cancers (Basel)*. 2021;13(6).
50. Fujiwara-Tani R, Sasaki T, Fujii K, Luo Y, Mori T, Kishi S, et al. Diabetes mellitus is associated with liver metastasis of colorectal cancer through production of biglycan-rich cancer stroma. *Oncotarget*. 2020;11(31):2982-94.

51. Manupati K, Paul R, Hao M, Haas M, Bian ZC, Holm TM, et al. Biglycan Promotes Cancer Stem Cell Properties, NF κ B Signaling and Metastatic Potential in Breast Cancer Cells. *Cancers (Basel)*. 2022;14(2).
52. Karamanou K, Franchi M, Proult I, Rivet R, Vynios D, Brézillon S. Lumican Inhibits In Vivo Melanoma Metastasis by Altering Matrix-Effectors and Invadopodia Markers. *Cells*. 2021;10(4).
53. Karamanou K, Franchi M, Piperigkou Z, Perreau C, Maquart FX, Vynios DH, et al. Lumican effectively regulates the estrogen receptors-associated functional properties of breast cancer cells, expression of matrix effectors and epithelial-to-mesenchymal transition. *Sci Rep*. 2017;7:45138.
54. Hu X, Villodre ES, Larson R, Rahal OM, Wang X, Gong Y, et al. Decorin-mediated suppression of tumorigenesis, invasion, and metastasis in inflammatory breast cancer. *Commun Biol*. 2021;4(1):72.
55. Mao L, Yang J, Yue J, Chen Y, Zhou H, Fan D, et al. Decorin deficiency promotes epithelial-mesenchymal transition and colon cancer metastasis. *Matrix Biol*. 2021;95:1-14.
56. David G, Bai XM, Van der Schueren B, Marynen P, Cassiman JJ, Van den Berghe H. Spatial and temporal changes in the expression of fibroglycan (syndecan-2) during mouse embryonic development. *Development*. 1993;119(3):841-54.
57. Hua R, Yu J, Yan X, Ni Q, Zhi X, Li X, et al. Syndecan-2 in colorectal cancer plays oncogenic role via epithelial-mesenchymal transition and MAPK pathway. *Biomed Pharmacother*. 2020;121:109630.
58. Contreras HR, Fabre M, Granés F, Casaroli-Marano R, Rocamora N, Herreros AG, et al. Syndecan-2 expression in colorectal cancer-derived HT-29 M6 epithelial cells induces a migratory phenotype. *Biochem Biophys Res Commun*. 2001;286(4):742-51.
59. Teng YH, Aquino RS, Park PW. Molecular functions of syndecan-1 in disease. *Matrix Biol*. 2012;31(1):3-16.
60. Toba-Ichihashi Y, Yamaoka T, Ohmori T, Ohba M. Up-regulation of Syndecan-4 contributes to TGF- β 1-induced epithelial to mesenchymal transition in lung adenocarcinoma A549 cells. *Biochem Biophys Rep*. 2016;5:1-7.
61. Chen LL, Gao GX, Shen FX, Chen X, Gong XH, Wu WJ. SDC4 Gene Silencing Favors Human Papillary Thyroid Carcinoma Cell Apoptosis and Inhibits Epithelial Mesenchymal Transition via Wnt/ β -Catenin Pathway. *Mol Cells*. 2018;41(9):853-67.
62. Farfán N, Ocarez N, Castellón EA, Mejía N, de Herreros AG, Contreras HR. The transcriptional factor ZEB1 represses Syndecan 1 expression in prostate cancer. *Sci Rep*. 2018;8(1):11467.
63. Poblete CE, Fulla J, Gallardo M, Muñoz V, Castellón EA, Gallegos I, et al. Increased SNAIL expression and low syndecan levels are associated with high Gleason grade in prostate cancer. *Int J Oncol*. 2014;44(3):647-54.
64. Reinhold MI, Kapadia RM, Liao Z, Naski MC. The Wnt-inducible transcription factor Twist1 inhibits chondrogenesis. *J Biol Chem*. 2006;281(3):1381-8.
65. Seki K, Fujimori T, Savagner P, Hata A, Aikawa T, Ogata N, et al. Mouse Snail family transcription repressors regulate chondrocyte, extracellular matrix, type II collagen, and aggrecan. *J Biol Chem*. 2003;278(43):41862-70.

66. Zhang Y, Zou X, Qian W, Weng X, Zhang L, Zhang L, et al. Enhanced PAPSS2/VCAN sulfation axis is essential for Snail-mediated breast cancer cell migration and metastasis. *Cell Death Differ.* 2019;26(3):565-79.
67. Cui X, Wang Y, Lan W, Wang S, Cui Y, Zhang X, et al. SPOCK1 promotes metastasis in pancreatic cancer via NF- κ B-dependent epithelial-mesenchymal transition by interacting with I κ B- α . *Cell Oncol (Dordr).* 2022;45(1):69-84.
68. Cook DP, Vanderhyden BC. Context specificity of the EMT transcriptional response. *Nat Commun.* 2020;11(1):2142.
69. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell.* 2017;171(7):1611-24.e24.
70. Nieto MA. Context-specific roles of EMT programmes in cancer cell dissemination. *Nat Cell Biol.* 2017;19(5):416-8.
71. Noborn F, Nilsson J, Larson G. Site-specific glycosylation of proteoglycans: A revisited frontier in proteoglycan research. *Matrix Biol.* 2022;111:289-306.
72. Zappia J, Joiret M, Sanchez C, Lambert C, Geris L, Muller M, et al. From Translation to Protein Degradation as Mechanisms for Regulating Biological Functions: A Review on the SLRP Family in Skeletal Tissues. *Biomolecules.* 2020;10(1).
73. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics.* 2016;32(18):2847-9.
74. Brock G, Pihur V, Datta S, Datta S. cIValid: An R Package for Cluster Validation. *Journal of Statistical Software.* 2008;25(4):1 - 22.
75. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol.* 2024;42(2):293-304.
76. Morabito S, Reese F, Rahimzadeh N, Miyoshi E, Swarup V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Rep Methods.* 2023;3(6):100498.
77. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell.* 2021;184(13):3573-87 e29.
78. Van den Berge K, Roux de Bezieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat Commun.* 2020;11(1):1201.
79. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford).* 2017;2017.

Figures

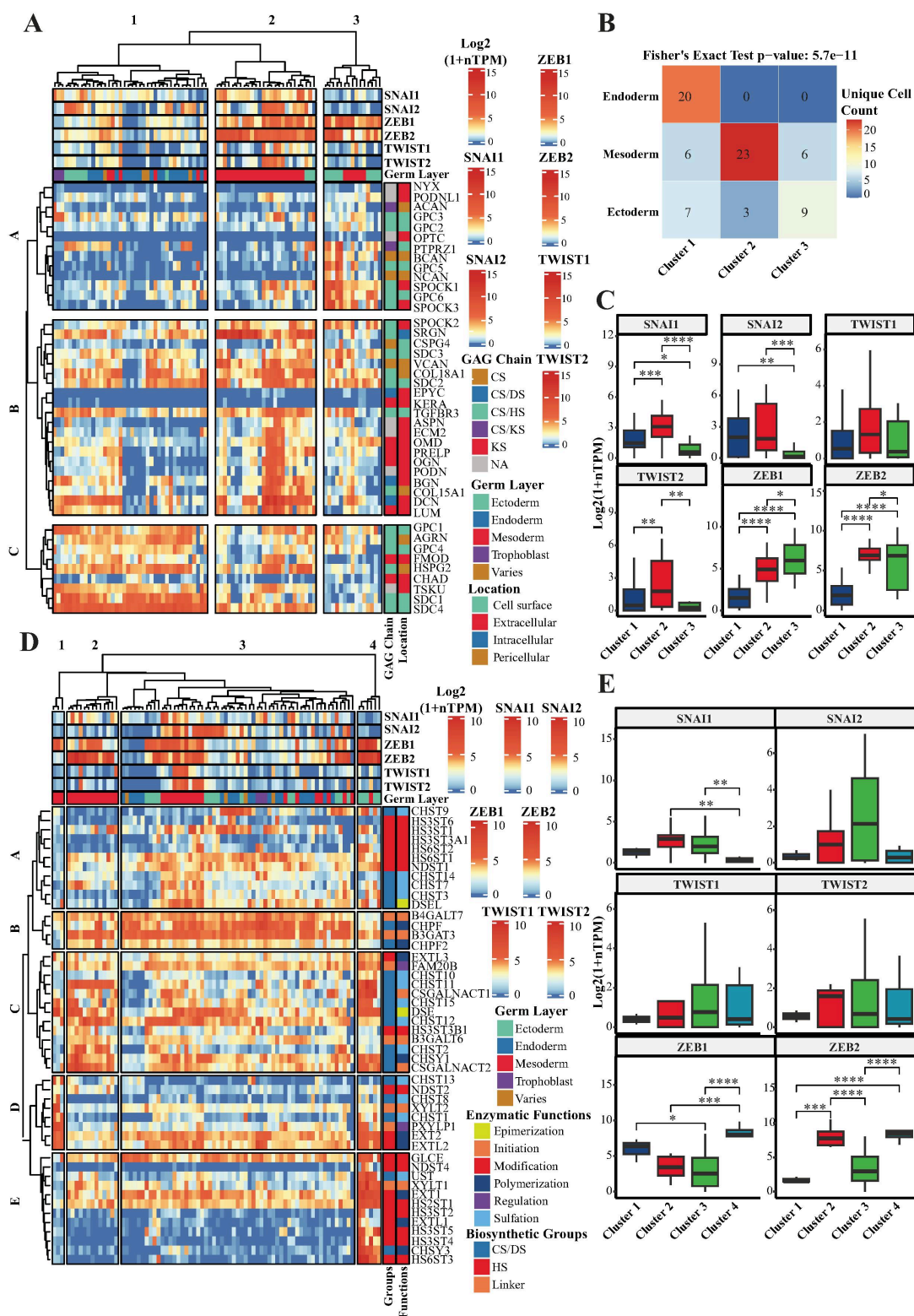


Figure 1

Proteoglycan core proteins, GAG biosynthetic enzymes, and EMT transcription factor expression.

Hierarchical clustering of adult human cell types based on proteoglycan core protein expression revealed clusters enriched for distinct embryonic germ layers, alongside corresponding patterns of EMT transcription factor (TF) expression. In contrast, clustering based on GAG biosynthetic enzyme expression showed predominantly tissue- and cell type-specific organization. **(A)** Heatmap showing

hierarchical clustering of proteoglycan core proteins across 81 normal human cell types (HPA), using $\log_2$ (nTPM+1) expression values. Top annotations show germ-layer origin (endoderm, mesoderm, ectoderm) and expression levels of canonical EMT TFs for each cell type. The right-side annotation bar indicates the GAG chain types associated with each proteoglycan. **(B)** Tile plot summarizing the number of cell types from each germ layer (endoderm, mesoderm, ectoderm) present within each proteoglycan core protein based cluster identified in **(A)**. **(C)** Boxplots representing the $\log_2$ (nTPM+1) expression values of the EMT TFs in each of the heatmap clusters as seen in **(A)**. **(D)** Heatmap showing hierarchical clustering of GAG- biosynthetic enzymes across the same cell types, using $\log_2$ (nTPM +1) expression values. Top annotations match those in **(A)**. The right-side bar represents the biosynthetic groups and function of each GAG biosynthetic enzyme. **(E)** Boxplots showing the $\log_2$ (nTPM+1) expression values of the EMT TFs in each of the heatmap clusters as seen in **(D)**.

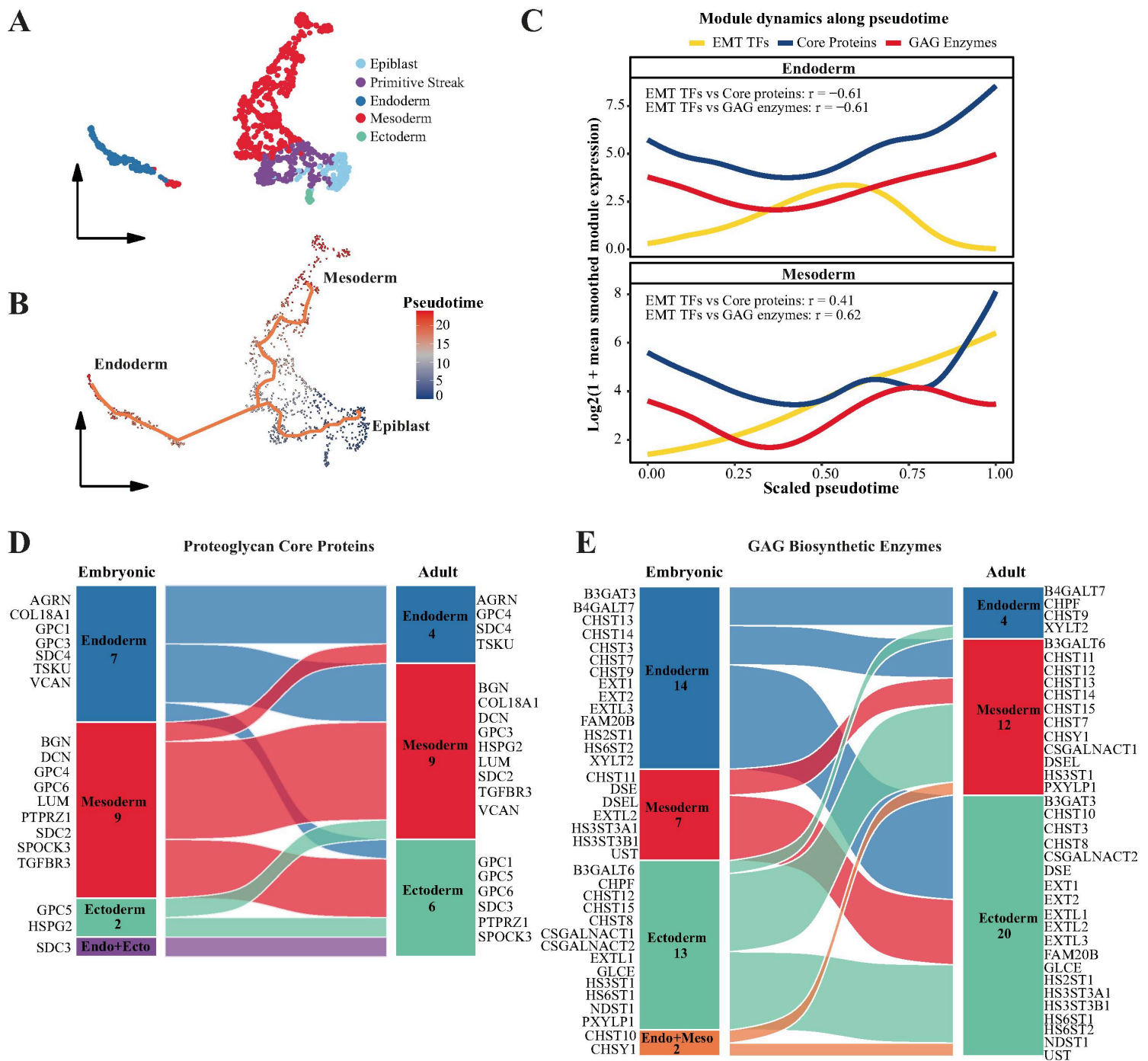


Figure 2

Embryonic dynamics and adult lineage conservation of proteoglycan core proteins and GAG biosynthetic enzymes.

Pseudotime analysis revealed coordinated expression dynamics of proteoglycan core proteins, GAG biosynthetic enzymes, and EMT transcription factors (TFs) during gastrulation and enabled comparison of embryonic and adult germ-layer specificity **(A)** UMAP embedding of single-cell transcriptomes from human gastrulation data, showing major developmental lineages: epiblast, primitive streak, endoderm, and mesoderm. **(B)** Diffusion pseudotime analysis illustrating developmental progression from epiblast toward endoderm and mesoderm. Cells are colored by pseudotime, with red colors indicating more differentiated states. **(C)** Scatterplots showing the relationship between EMT TF

module scores (yellow) and either the proteoglycan core protein module (blue) or GAG biosynthetic enzyme module (red), analyzed separately for endoderm (top) and mesoderm (bottom) trajectories. Curves represent smoothed pseudotime trends, and Pearson correlation coefficients (r) with corresponding p-values are displayed. All scores are derived from $\log_2(\text{Expression}+1)$ values. **(D-E)** Alluvial plots showing how proteoglycan core proteins **(D)** and GAG biosynthetic enzyme genes **(E)** are expressed across embryonic lineages (endoderm, mesoderm, ectoderm, endoderm+ectoderm (endo+ecto) and endoderm+mesoderm (endo+meso)) and adult tissue-derived cell types, using Human Protein Atlas cell-type annotations. Each plot illustrates how expression patterns observed during early embryonic development map onto cell identities found in adult tissues.

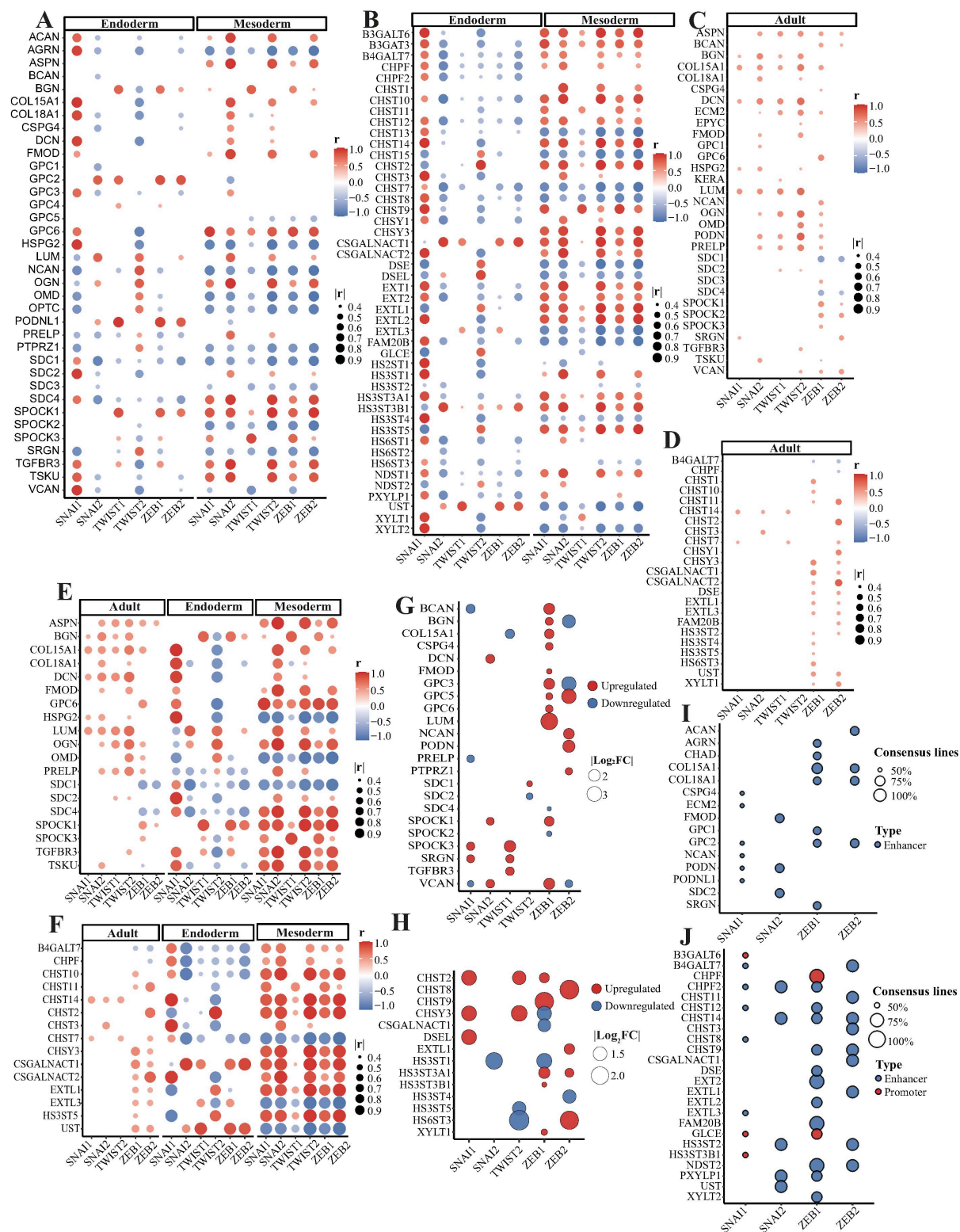


Figure 3

EMT transcription factors associations with proteoglycan core proteins and GAG biosynthetic enzymes. Correlation, perturbation, and ChIP-seq analyses identify proteoglycan core proteins and GAG biosynthetic enzymes that are candidate targets of EMT transcription factor (TF) regulation. **(A-B)** Pairwise correlations between EMT TF expression and proteoglycan core protein genes **(A)** or GAG biosynthetic enzyme genes **(B)** across endoderm and mesoderm gastrulation lineages. Only correlations

with $|r| > 0.4$ are shown. Bubble size represents correlation magnitude; color indicates sign (red = positive, blue = negative). **(C-D)** Same correlation analysis performed in adult tissue-derived cell types, showing proteoglycan core protein **(C)** and GAG biosynthetic enzyme genes **(D)** with $|r| > 0.4$ relative to EMT TF expression patterns. **(E-F)** Summary of proteoglycan core protein **(E)** and GAG biosynthetic enzyme **(F)** genes that meet the correlation threshold ($|r| > 0.4$) in adult, endoderm, or mesoderm contexts. **(G-H)** RNA-seq perturbation analyses showing proteoglycan core protein **(G)** and GAG biosynthetic enzyme **(H)** genes with consistent EMT TF-dependent regulation. Genes plotted have $Q3(|\log FC|) > 1$ across perturbation datasets. Bubble size reflects the number of supporting datasets; color denotes direction of regulation. **(I-J)** Consensus ChIP-seq regulatory interactions between EMT TFs and proteoglycan core protein **(I)** or GAG biosynthetic enzyme **(J)** genes. Bubble size indicates the fraction of datasets supporting the TF-gene assignment; color denotes enhancer-like vs promoter-like regulatory context.

Supplementary Files

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

- [Additionalfile3TableS6S11.xlsx](#)
- [Additionalfile6TableS16.xlsx](#)
- [Additionalfile2TableS1S5.xlsx](#)
- [Additionalfile4TableS12.xlsx](#)
- [Additionalfile5TableS13S15.xlsx](#)
- [Additionalfile1FigureS.pdf](#)