anexVis: Visual analytics framework for analysis of RNA expression Supplementary Materials

Diem-Trang Tran, Tian Zhang, Ryan Stutsman, Matthew Might,
Umesh R. Desai and Balagurunathan Kuberan

December 13, 2017

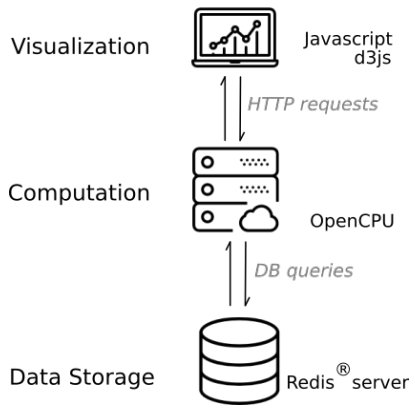
Contents

1 User guide

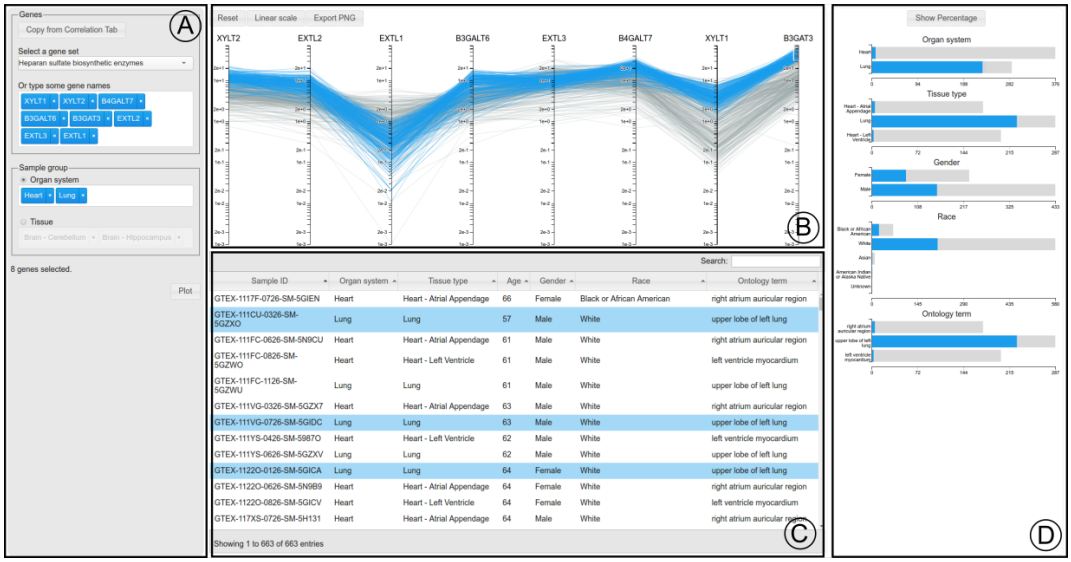
The implementation scheme is depicted in Supp Fig S1. In most cases, users are not expected to install and host the application unless there is a need to customize the data, for example, adding genes not available in the current version or adopting the framework for alternative data sets. For such needs, installation instruction and example script are provided at https://github.com/anexvis/setup.

The web application is accessible at https://anexvis.chpc.utah.edu/

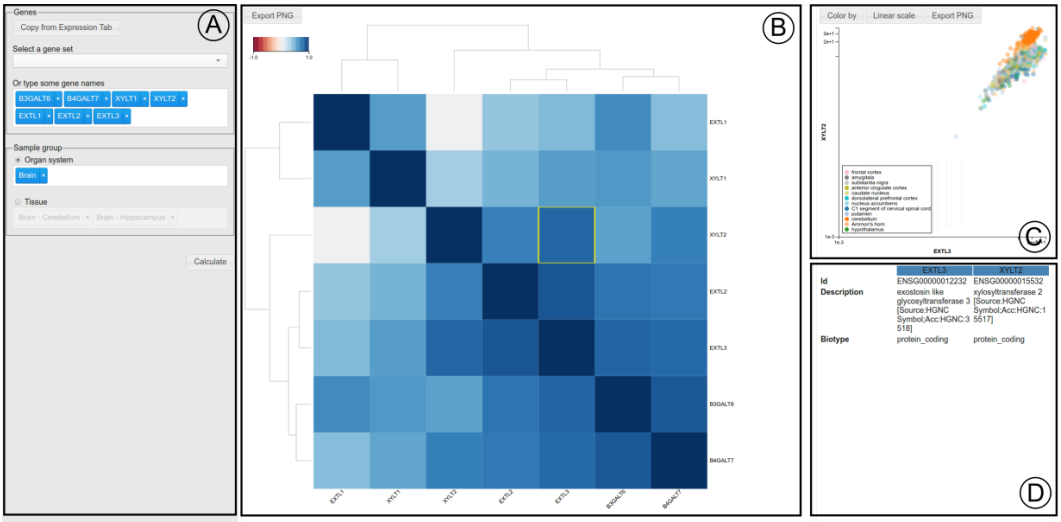
A tutorial video is provided online at https://youtu.be/IBQiUUsXJls. An equivalent description is given along with the screen-shots in Supp Fig S2 and Supp Fig S3.



Supp Fig. S1: **Architecture of the visualization framework**

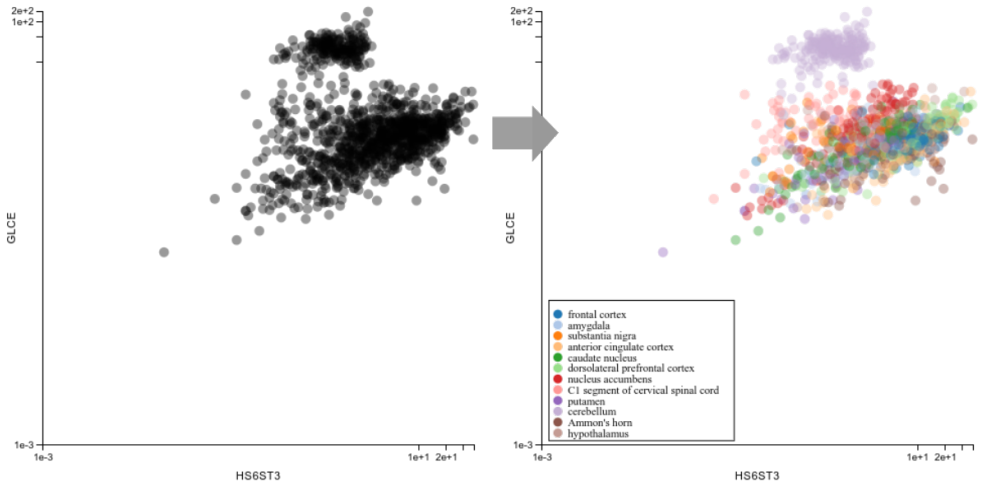


Supp Fig. S2: **Gene-based expression profile**. To visualize gene-based expression profiles , a user would start at the panel **A** by specifying the genes of interest, as well as the organ system(s) or tissue type(s) to include. Panel **B** will visualize the expression profile as parallel coordinate plot, as explained in the text. Panel **C** provides the detailed meta data of each samples included in the analysis. Panel **D** summarizes the fraction of highlighted samples in various perspectives.



Supp Fig. S3: **Correlation-based expression profile**. To visualize correlation-based expression profiles (Supp Fig S3), a user would start at panel **A** to specify the genes of interest, as well as the organ system(s) or tissue type(s) to be included. Panel **B** will display the correlation matrix among the selected genes, each square color-coding the correlation coefficient. Clicking on each square will give the corresponding scatter diagram in panel **C**. A description of the gene pair in query is shown in panel **D** for convenient reference of gene names and functions.

Supp Fig. S4: **Technical limits of RNA-seq experiments become more transparent**. Many genes are found to segregate the samples into populations of null expression (in blue) and of low expression (in grey). Using *anexVis*, one can quickly examine the composition of the null-expression population. In the case of HS3ST6 highlighted here, the null-expression population has equivalent representation in all histological sites, as well as genders and races, implying that HS3ST6 transcript abundance in some of the samples was too low to be detected.



Supp Fig. S5: **Coloring of scatter plot by sample attributes reveals meaningful structures**. An example pairwise co-expression pattern in brain samples, where correlation coefficient is low (-0.015), yet the pattern is intriguing. Coloring the data points by ontology term reveals that the cerebellum is distinct from the rest of brain regions.

2 Comparison with other RNA-seq visualization tools

There has been an active development of web-based tools for visualization of data generated at various stages in an RNA-seq workflow. Specifically, *QuickRNASeq* (Zhao *et al.*, 2016) simplifies the workflow by pipelining the available processing and analysis tools into user-friendly graphical interface. Its plethora of plots to explore QC metrics of RNA-seq samples also enable rigorous QC after read mapping. *ASAP* (Gardeux *et al.*, 2017) has a similar intention yet focuses on the later steps such as differential expression analysis, dimensionality reduction, etc. *DEIVA* (Harshbarger *et al.*, 2017) leverages traditional differential expression (DE) analysis with graphical user interface, interactive MA plot and Volcano plot. *Shinyheatmap* (Khomtchouk *et al.*, 2017) and *NG-CHM* (Broom *et al.*, 2017) improve the heatmap representation of gene expression data, in terms of performance and customization, respectively. *anexVis*, introduced by this manuscript, furthers gene expression analysis by integrating the different views on multiple data types, including gene expression, sample metadata, and phenotypes. The integrated views supported by anexVis are particularly helpful in exploring the multi-gene expression and co-expression patterns across different tissue types. Together, these tools contribute to a variety of options for visualizing transcriptomic data.

Depending on the biological questions and correspondingly, the required analytic tasks, visualization solutions may vary significantly. Thus, the choice of an appropriate tool from a user viewpoint should also be based primarily on the questions one wants to address. The plotting features such as zooming and panning, generating publication images, customizing plot parameters, etc., although important, have become increasingly universal and thus, secondary in deciding which tool is appropriate. In such perspective, we provide a summary of these tools in Table S1.

	QC metrics on samples	Expression levels	Differential expression	Expression patterns	Co-expression patterns	Expression patterns vs Phenotypes	Performance of co-expression measure
QuickRNASeq	++						
DEIVA		+	++				
Shinyheatmap		+	+	+	+		
ASAP		+	+	+	+		
NG-CHM		+	+	+	+	+	
anexVis		++	+	++	++	++	++

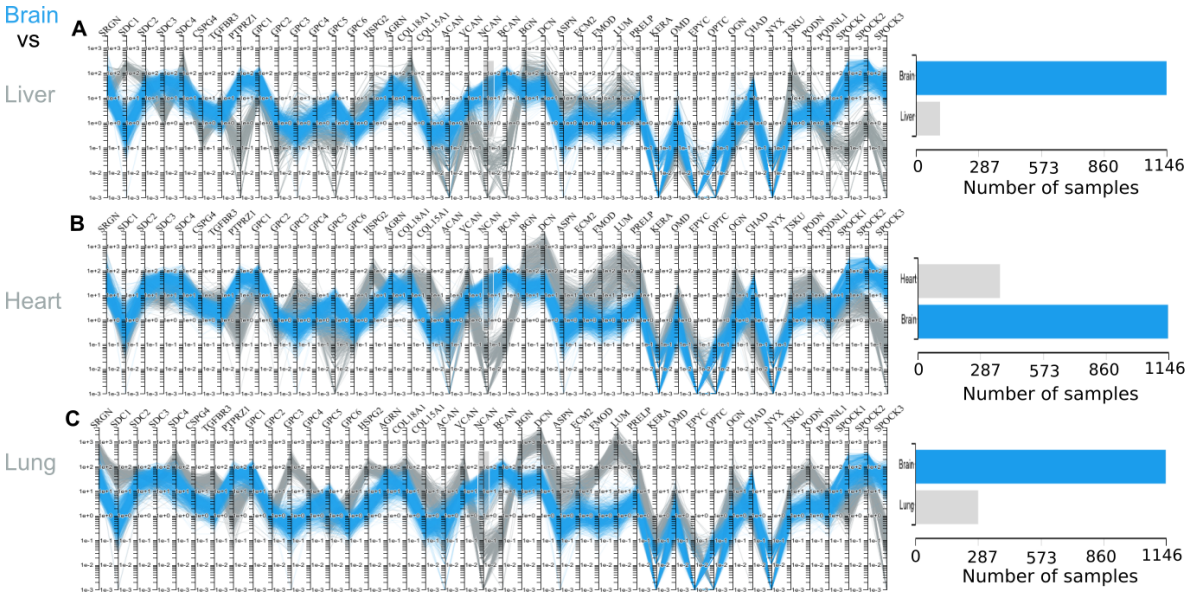
Table S1: **Summary of recent RNA-seq visualization tools, by the level of support for each analytic task.** The higher level of support indicates a more appropriate tool for the corresponding purpose.

[blank]	Not supported
+	Data can be read but difficult to compare, constrast, or relate
++	Data are presented to highlight an insight

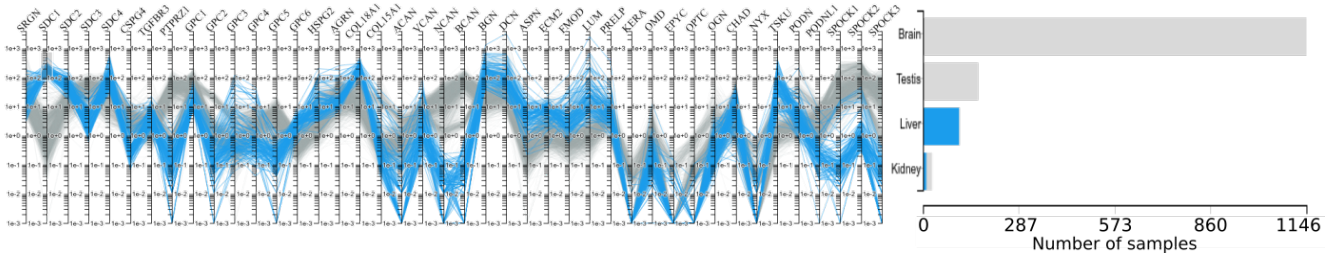
3 Example use cases

3.1 Example 1: Tissue signatures by proteoglycan core proteins

Proteoglycan structural diversity is attributed to both the core proteins and the attached glycosaminoglycan chains. More than 40 core proteins have thus far been identified (Iozzo and Schaefer, 2015) and the tissue distribution of many of these proteins have been established. The tissue-specific expression patterns can be readily detected with our interactive visualization of the expression data. BCAN (brevican) and NCAN (neurocan) are known to be brain-specific proteoglycans (Rauch *et al.*, 1991; Yamada *et al.*, 1994; Frischknecht and Seidenbecher, 2012). By plotting samples from the brain, liver, heart and lung, one can



Supp Fig. S6: **Comparison of transcriptional profiles of proteoglycan core proteins in various tissue pairs.** (A) Brain vs Liver. (B) Brain vs Heart (C) Brain vs Lung. In all plots, brain samples are highlighted in blue. Some brain-specific proteoglycans can be readily recognized such as neurocan NCAN and brevican BCAN. Many others also form tissue-characteristic cluster, such as fibromodulin (FMOD), syndecan-1 (SDC1), glypican-3 (GPC3).



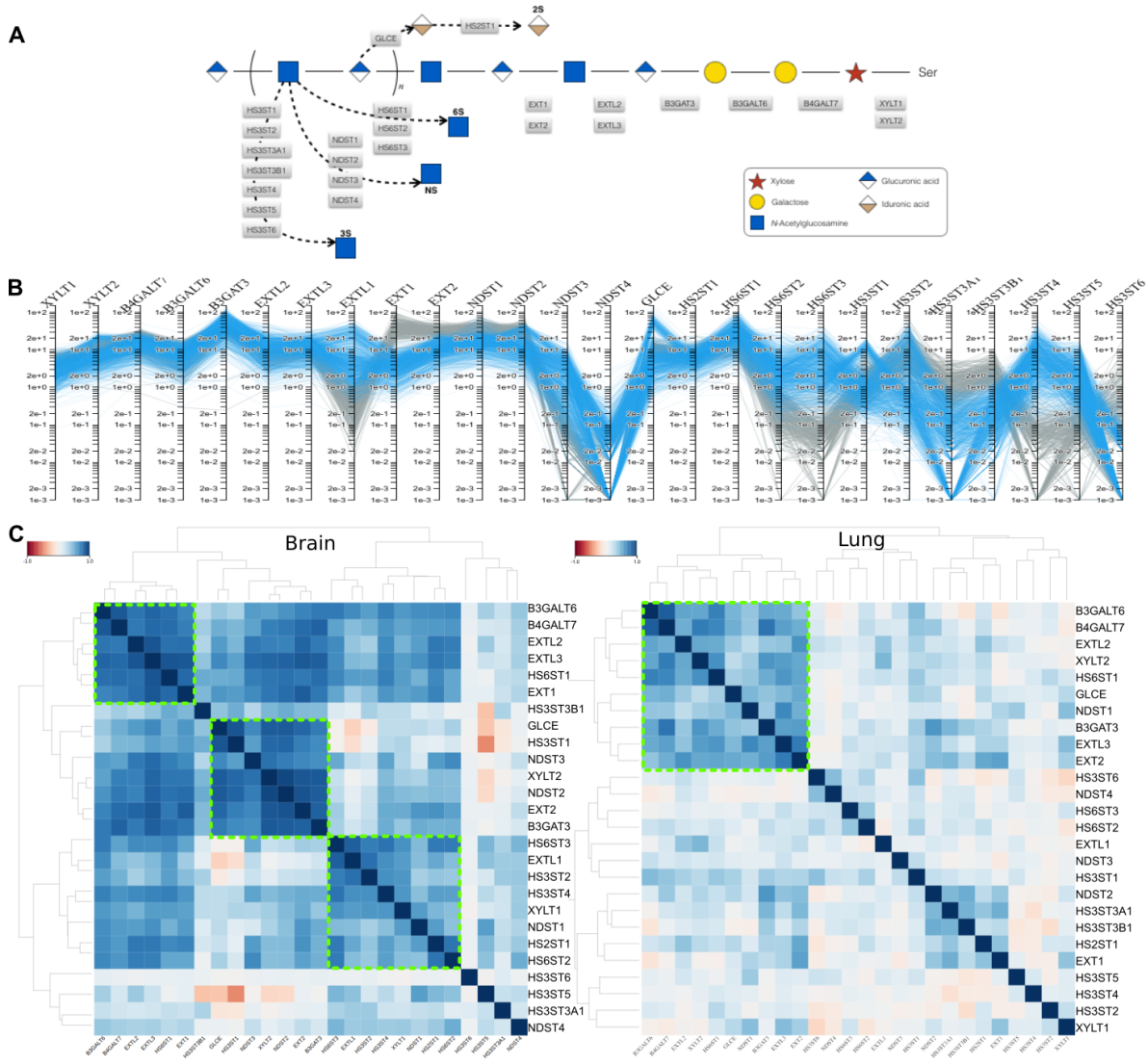
Supp Fig. S7: The expression profiles of proteoglycan core proteins in human revealed similar tissue distribution as in mouse tissues investigated by *Nairn et al.* (2008) . For example, SDC1 (syndecan-1) is highly expressed in liver, while BGN (biglycan) is similary expressed in all four tissues.

visually recognize the distinct clusters on these two axes and many others. Highlighting the samples with high expression of NCAN confirms that they are exclusively of brain origin (Fig S6). Several other axes also provide interesting insights. For example, FMOD (fibromodulin) expression is distinctively high in lung and heart, SDC1 (syndecan-1) is highly expressed in lung and liver, GPC3 (glypican-3) is highly expressed in lung and some samples of the right atrial appendage of heart. Together, these proteoglycan genes compose expression profiles that uniquely represent each tissue type.

Compared with transcriptional profiles of mouse tissues (Nairn *et al.*, 2008), many similar features are also observed in human; for example, the distinctively high level of SDC1 (syndecan-1) in liver and the universal expression of BGN (biglycan) in all four tissues (Supp Fig S7).

3.2 Example 2: Tissue signatures by heparan sulfate biosynthetic genes

Heparan sulfate (HS) is composed of repeating disaccharide units of glucosamine and hexuronic acid. A variety of modifications on HS chains, including *C*-5-epimerization and 2-*O*-sulfation of the hexuronic acid, *N*-deacetylation and *N*-sulfation, 6-*O*-sulfation and 3-*O*-sulfation of glucosamine, enable HS to be an information-rich molecule that could render binding specificity to a wide range of protein ligands which in turn regulate cellular signaling events. Analysis of tissue-specific HS structures has been of great interest for over two decades. Early biochemical studies on various biological systems have confirmed the



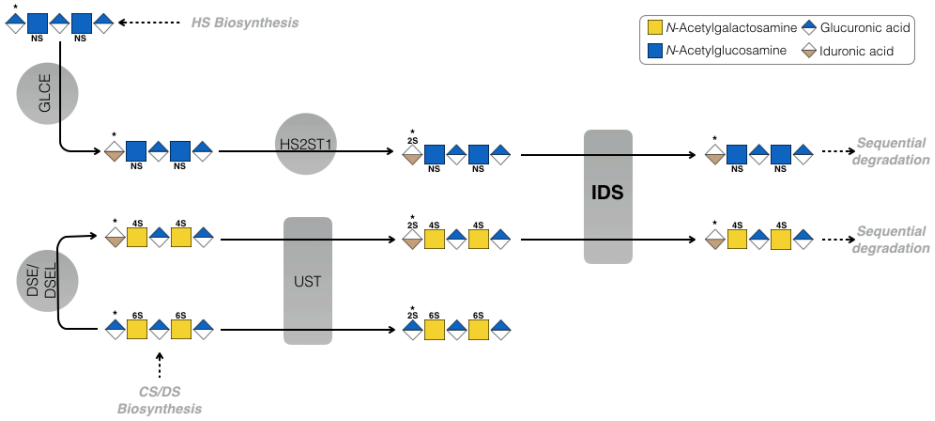
Supp Fig. S8: **A**. Heparan sulfate biosynthetic pathway and the corresponding genes. **B**. Expression profiles of HS biosynthetic enzymes in brain (blue) and lung (grey). **C**. The pairwise correlation among HS biosynthetic genes in brain and lung reveals distinct co-expression patterns. Clusters of highly co-expressed genes are highlighted in green boxes.

presence of organ-specific HS structural profiles in animals including mouse (Ledin *et al.*, 2004), rat (Shi and Zaia, 2009), and bovine (Shao *et al.*, 2013). However, it remains unknown how these signatures are generated. The process of generating HS fine structures involves inscrutable coordination among a series of biosynthetic enzymes numbering well over 25. A schematic description of this pathway is provided in Fig S8A. Each of these biochemical reactions is catalyzed by specific enzymes, most of which exists in multiple isoforms, i.e. encoded by multiple genes. It is reasonably hypothesized that the combinatorial expression of these enzyme-encoding genes dictates, at least partly, the tissue-specific HS structures.

Fig S8-B and -C reveal the distinct expression profiles and co-expression patterns, respectively, of HS biosynthetic genes in brain and lung. The expression profiles revealed that most of the linkage enzymes, and some modification enzymes (NDST1, NDST2, HS2ST1, HS6ST1) are similarly expressed, while most modification enzymes (NDST3, NDST4, HS6ST2, HS6ST3 and the HS3STs) vary widely, across the two organs and even within an organ. These observations of the human tissues align with previous observations of mouse tissues (Nairn *et al.*, 2008). In addition, the co-expression patterns of these genes revealed interesting patterns. Many of them are highly co-expressed in brains, forming tightly correlated clusters, such as those of (B3GALT6, B4GALT7, EXTL2, EXTL3, EXT1, HS6ST1) or of

(GLCE, HS3ST1, NDST3, XYLT2, NDST2, EXT2, B3GAT3). The few genes whose expression does not couple tightly with any of the other genes are the 3-*O*-sulfotransferases, HS3ST3A1, HS3ST5, HS3ST6, and an *N*-deacetylase/*N*-sulfotransferase, NDST4. In contrast, most of these genes are not coupled in lung. Here the co-expression of 3-*O*-sulfotransferase with other genes, if present, is usually weak. This comparison suggests that the HS structures in brain are bestowed with more intricate, and likely more strictly regulated modifications.

3.3 Example 3: Understanding congenital disorders of glycosylation



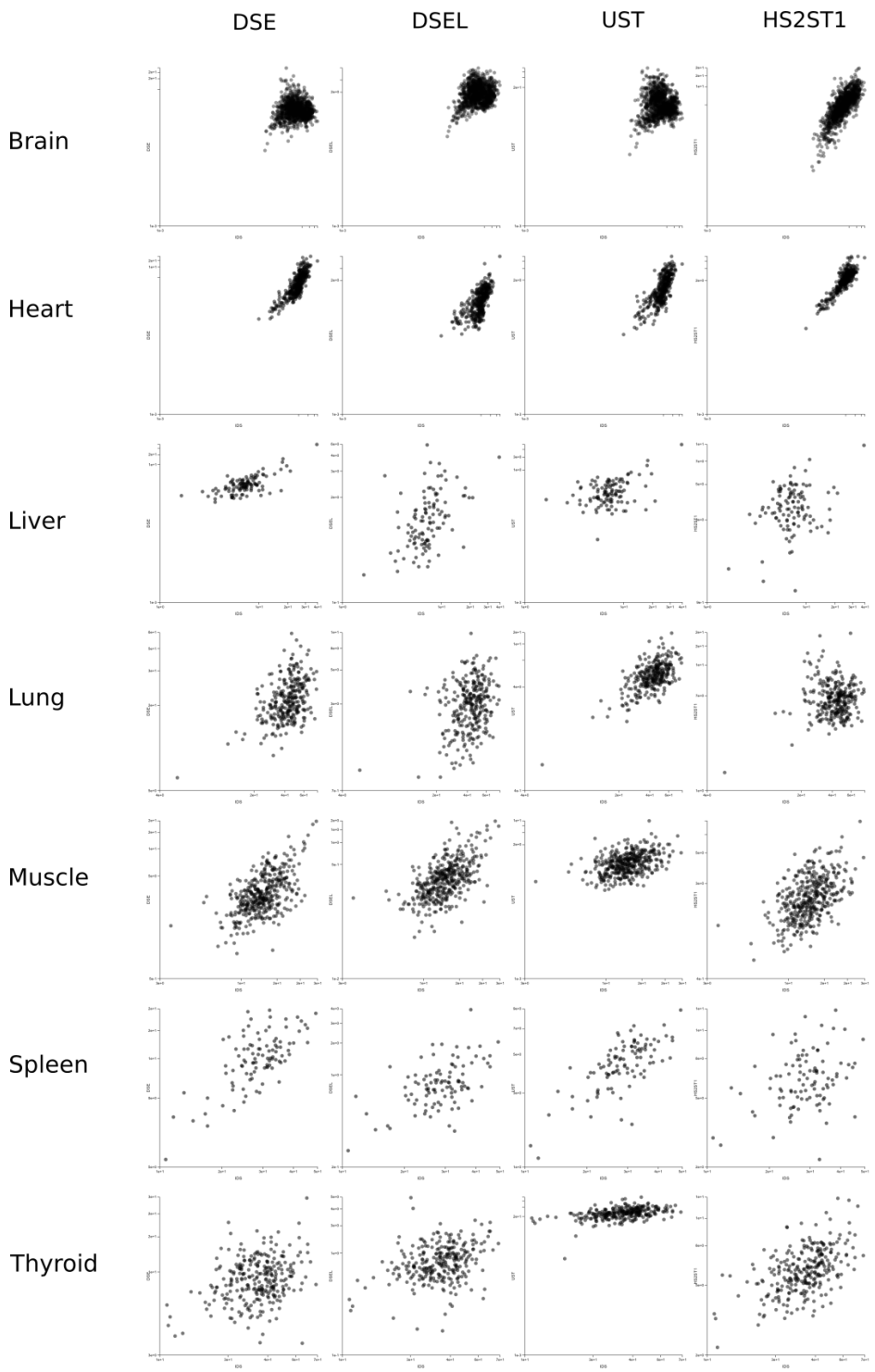
Supp Fig. S9: **IDS-related reactions**. Iduronate sulfatase, IDS, catalyzes the removal of 2-*O*-sulfate groups on the iduronic acid of either HS or DS. Thus the substrates of IDS come from two distinct pathways: (1) HS biosynthesis, and (2) CS/DS biosynthesis. HS substrate of IDS is generated via the conversion of glucuronic acid to iduronic acid (epimerization) by GLCE, and then 2-*O*-sulfation by HS2ST1. Similarly, DS substrate of IDS is generated via epimerization by DSE/DSEL and then 2-*O*-sulfation by UST. Unlike HS2ST1, which highly prefers iduronic acid over glucuronic acid as a substrate, UST can modify either glucuronic acid or iduronic acid. Thus the activity of HS2ST1 is enough to indicate the presence of HS substrate, while the activity of both DSE/DSEL and UST are needed to indicate the presence of DS substrate.

A large number of congenital glycosylation diseases that have been documented until now are the result of deficiency in one of the glycan metabolic activities (Freeze and Schachter, 2009). In most cases, such defect introduces an imbalance between biosynthesis and metabolism, resulting in abnormalities in multiple tissues. When the rescue of metabolic activity is difficult, the inhibition of biosynthesis could be helpful. Thus the knowledge of relative expression of the biosynthetic and metabolic genes is essential for devising therapeutic strategies.

One of such diseases is Hunter syndrome, which is caused by deficiency in the iduronyl sulfatase enzyme. This enzyme, encoded by the gene IDS, catalyzes the removal of the sulfate group on the iduronic acid unit of HS and dermatan sulfate (DS) (Fig S9). Hence, IDS deficiency prevents these GAGs from being degraded properly, resulting in the accumulation of HS and DS chains and disrupting the fine balance between synthesis and turnover. Assuming that abnormal phenotypes are caused by the deviation from this balance, we can then predict the affected tissues, based on the normal relation between IDS and the related biosynthetic genes. The generation of 2-*O*-sulfated iduronic acid on DS involves the activity of DSE/DSEL and UST, while that on HS involves HS2ST1 (Fig S9).

Additionally, assuming that transcriptional regulation accounts for the large part of GAG structures, we would expect that a tissue synthesizing 2-*O*-sulfated DS to exhibit co-expression of DSE/DSEL and UST. Hypothetically if IDS is found to correlate with both of these genes, the balance of DS synthesis and turnover in such places is tightly regulated and a disruption in either direction would be detrimental. Similarly, if IDS is highly correlated with HS2ST1, the balance of HS synthesis and turnover is tightly regulated. To predict whether IDS deficiency would affect a given tissue, we qualitatively examined the co-expression pattern of IDS with each of these genes (Supp Fig S10).

It should be noticed that 2-*O*-sulfation by HS2ST1 almost always requires epimerization, i.e. conver-



Supp Fig. S10: Pairwise correlation in transcriptional abundance of IDS and 2-*O*-sulfated iduronate generating enzymes: DSE/DSEL, UST, HS2ST1.

sion of glucuronic acid to iduronic acid in the HS chain, while the 2-*O*-sulfation by UST can act on either glucuronic acid or iduronic acid (Silbert and Sugumaran, 2002; Mikami and Kitagawa, 2013). Correspondingly, the coexpression of IDS with HS2ST1 is enough to indicate the coupling between biosynthesis and metabolism of HS-2-*O*-sulfated iduronic acid, while the conjunctive activity of UST and DSE/DSEL is necessary to indicate the presence of DS-2-*O*-sulfated iduronic acid. A boolean encoding is used to

Table S2: **Co-expression of IDS and related biosynthetic genes.** The pairwise co-expression is binary-encoded with 0 for no co-expression, and 1 for detectable co-expression.

	Co-expression with IDS					
	DSE	DSEL	UST	HS2ST1	Predicted effect	Relevant phenotype
Boolean variable	A	B	C	D	$(A \wedge C) \vee (B \wedge C) \vee D$	
Lung	1	0	0	0	0	
Liver	1	0	0	0	0	enlargement
Brain	0	0	0	1	1	intellectual disability, (jerky movements)
Heart	1	1	1	1	1	leaky heart valves
Thyroid	0	0	0	0	0	
Muscle	1	1	1	1	1	jerky movements
Spleen	1	0	1	0	1	enlargement

denote this relation: 1 if there is strong co-expression, 0 otherwise. Using our tool, the pairwise coupling of the four biosynthetic genes with IDS can be filled as shown in Table S2. This table suggests that brain, heart, muscle and spleen are affected, while lung, liver and thyroid are not. These sites of effects closely match with documented phenotypes in Hunter syndrome: intellectual disability, leaky heart valves, enlarged spleen, and jerky movements (Haldeman-Englert, 2015). The fact that liver was not predicted to be impaired in IDS deficiency may imply that liver enlargement observed in this syndrome, as in many other disorders, could be a secondary effect. In general, our visual analysis on RNA expression data, with the aid of this framework, was able to pinpoint the affected organs in IDS deficiency. Such analysis illustrates how visualization can provide preliminary ideas to develop quantitative models of human disease conditions.

References

Broom, B. M., Ryan, M. C., Brown, R. E., *et al.* (2017). A Galaxy Implementation of Next-Generation Clustered Heatmaps for Interactive Exploration of Molecular Profiling Data. *Cancer Research*, **77**(21), e23–e26.

Freeze, H. H. and Schachter, H. (2009). Genetic Disorders of Glycosylation. In A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart, and M. E. Etzler, editors, *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2nd edition.

Frischknecht, R. and Seidenbecher, C. I. (2012). Brevican: A key proteoglycan in the perisynaptic extracellular matrix of the brain. *The International Journal of Biochemistry & Cell Biology*, **44**(7), 1051–1054.

Gardeux, V., David, F. P. A., Shajkofci, A., Schwalie, P. C., and Deplancke, B. (2017). ASAP: A web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics*, **33**(19), 3123–3125.

Haldeman-Englert, C. (2015). Hunter syndrome. In *A.D.A.M Medical Encyclopedia [Internet]*.

Harshbarger, J., Kratz, A., and Carninci, P. (2017). DEIVA: A web application for interactive visual analysis of differential gene expression profiles. *BMC Genomics*, **18**, 47.

Iozzo, R. V. and Schaefer, L. (2015). Proteoglycan form and function: A comprehensive nomenclature of proteoglycans. *Matrix Biology*, **42**, 11–55.

Khomtchouk, B. B., Hennessy, J. R., and Wahlestedt, C. (2017). Shinyheatmap: Ultra fast low memory heatmap web interface for big data genomics. *PLOS ONE*, **12**(5), e0176334.

Ledin, J., Staatz, W., Li, J.-P., *et al.* (2004). Heparan Sulfate Structure in Mice with Genetically Modified Heparan Sulfate Production. *Journal of Biological Chemistry*, **279**(41), 42732–42741.

Mikami, T. and Kitagawa, H. (2013). Biosynthesis and function of chondroitin sulfate. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1830**(10), 4719–4733.

Nairn, A. V., York, W. S., Harris, K., *et al.* (2008). Regulation of Glycan Structures in Animal Tissues TRANSCRIPT PROFILING OF GLYCAN-RELATED GENES. *Journal of Biological Chemistry*, **283**(25), 17298–17313.

Rauch, U., Gao, P., Janetzko, A., *et al.* (1991). Isolation and characterization of developmentally regulated chondroitin sulfate and chondroitin/keratan sulfate proteoglycans of brain identified with monoclonal antibodies. *Journal of Biological Chemistry*, **266**(22), 14785–14801.

Shao, C., Shi, X., Phillips, J. J., and Zaia, J. (2013). Mass spectral profiling of glycosaminoglycans from histological tissue surfaces. *Analytical Chemistry*, **85**(22), 10984–10991.

Shi, X. and Zaia, J. (2009). Organ-specific Heparan Sulfate Structural Phenotypes. *Journal of Biological Chemistry*, **284**(18), 11806–11814.

Silbert, J. E. and Sugumaran, G. (2002). Biosynthesis of Chondroitin/Dermatan Sulfate. *IUBMB Life*, **54**(4), 177–186.

Yamada, H., Watanabe, K., Shimonaka, M., and Yamaguchi, Y. (1994). Molecular cloning of brevican, a novel brain proteoglycan of the aggrecan/versican family. *Journal of Biological Chemistry*, **269**(13), 10119–10126.

Zhao, S., Xi, L., Quan, J., *et al.* (2016). QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization. *BMC Genomics*, **17**, 39.