GlyCompare

Bokan, Ben

Take home message

- Study the motif based on glycoprofile from the synthesis network perspective.
- The biosynthesis of glycan is determined by glycosyltransferases that assemble monosaccharides moieties into linear and branched glycan chains step by step . Since there is no template, the synthesis is totally stochastic process. A glycoprofile has multiple glycan and each glycan are the product of a complex network. It is very hard to simulate the synthesis of each glycan individually by using stochastic model. However, the glycoprofile provides a macroscopic description of this stochastic system. If we break down all glycan into substructures, we can study the progress of this complex network comprehensively.

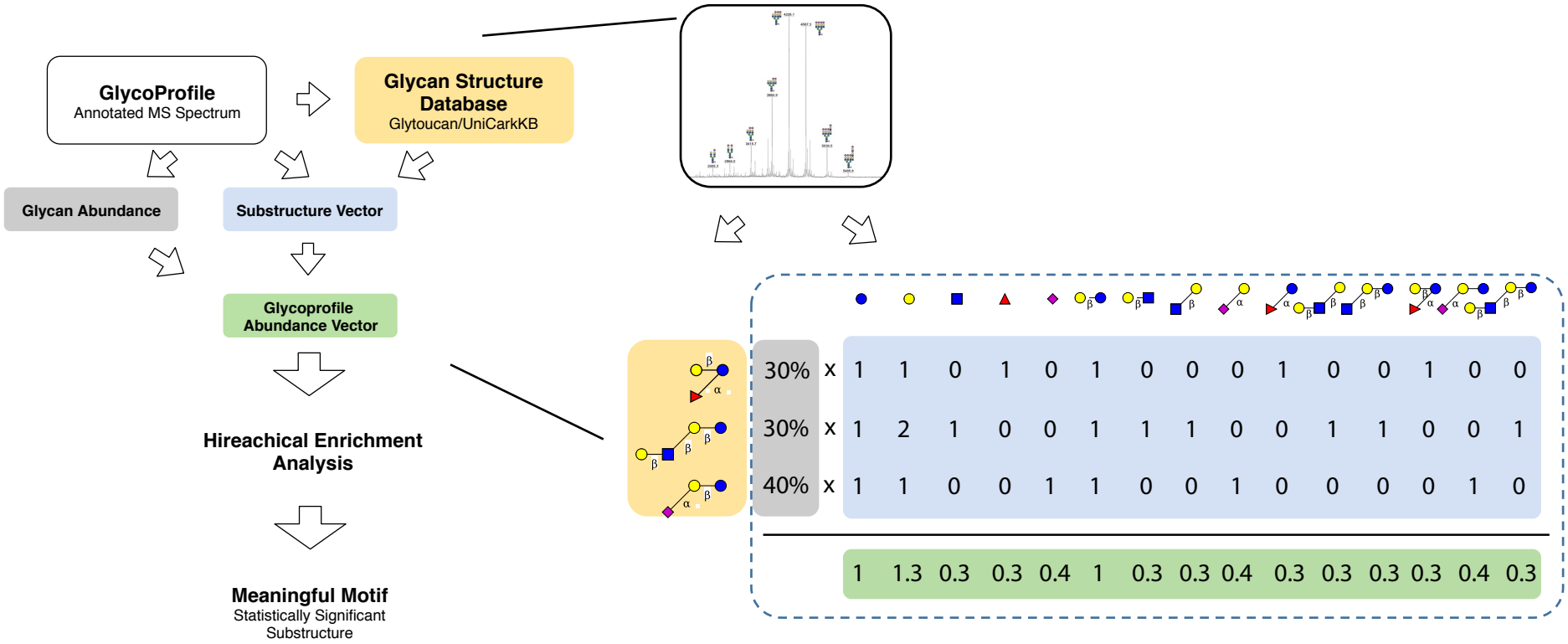
Compare the glycoprofile

- Comparing the glycoprofile is important. It tells us how glycan synthesis process is regulated or influenced by the genome modification. Difference, substructre(motif), synthesis perturbation.
- Reading glycoprofile is hard. Each glycoprofile has multiple glycan structures and each glycan structure has overlapping chemical structure.
- Comparing glycoprofile is even harder. Because each glycoprofile has different glycans, you cannot directly compare them.

GlyCompare

- Comparing large set of glycoprofiles
- Tracking the substructure(motif) abundance change
- Correlating the phenotypes with the motif’s abundance.
- Illustrating the perturbation of the synthesis netowork.

Figure 1. Workflow and idea of data preprocessing.

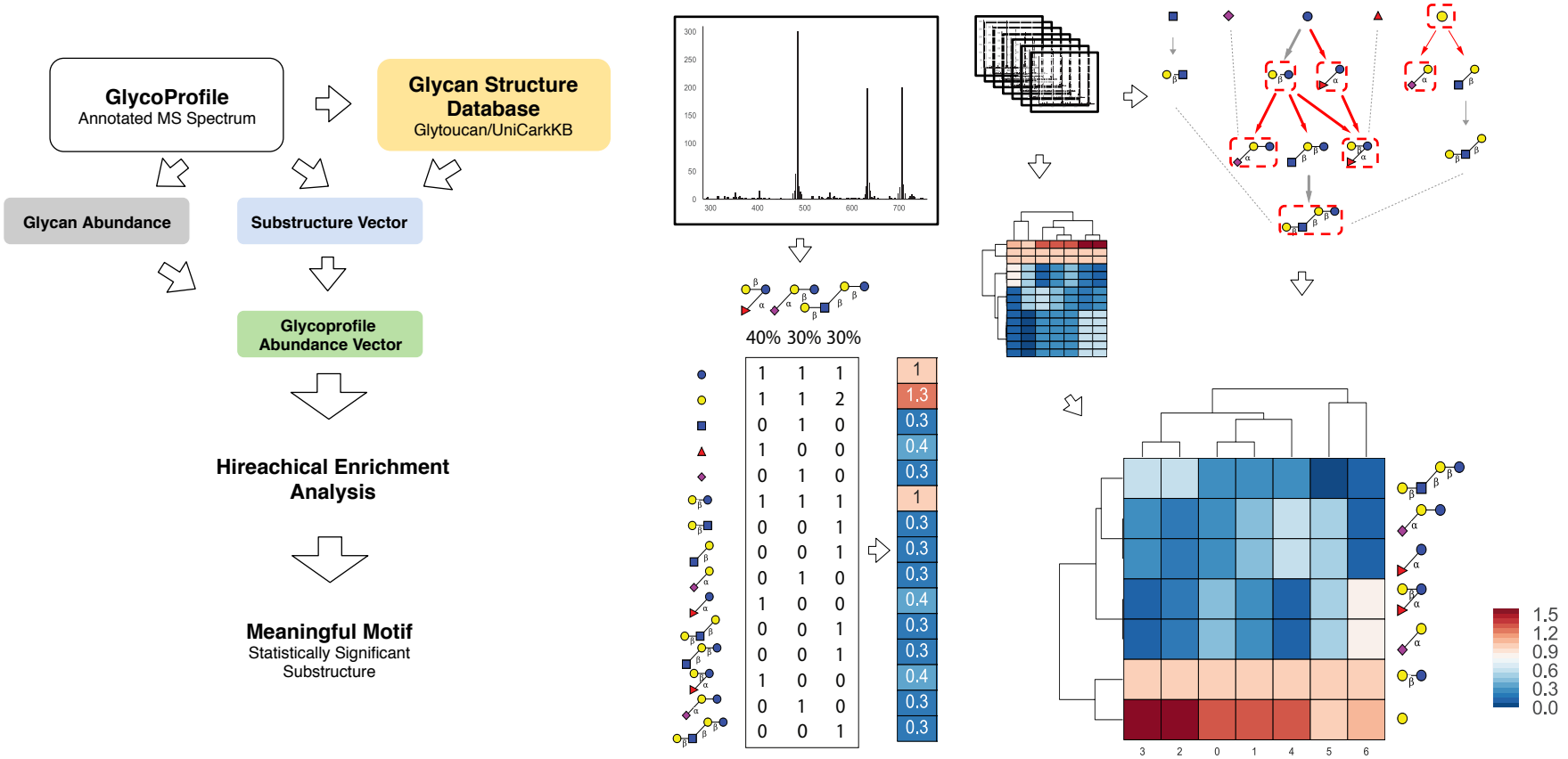


The workflow of the analysis pipeline. The data from Glycoprofile and Glycan structure are vectored to numeric vector and then go to the hierarchical analysis.

White box: pseudo glycoprofile
Orange box : glycan structure
Green box : glycoprofile abundance vector
Grey box : glycan abundance

This plot shows that how the glycoprofile data is transformed to glycan **Substructure Vector** and **Glycoprofile Abundance Vector**

Figure1 Working pipeline



2 Result

We demonstrate our analysis on 16 EPO glycoprofiles curated from NBT paper, the CHO cell lines includes the single or joint knockout from B4galT1/2/3/4/ family, st3gal4/6 family and mgat1/2/4A/4B/5 family. For each glycoprofile, the glycan structure and the relative abundance are obtained.

Result 2.1 Clustering of 16 glycoprofiles

Figure 2A
A glycoprofile vector with 722 glycan substructure are generated and then reduced to 118 glycan substructures. The cluster map (Figure2 A) shows there are 7 different types of glycoprofile. Each cluster has distinguished glycan structure patterns and the distance between each cluster demonstrate the structure difference between each groups of knockout. The cluster contains WT also has B4GalT1, B4GalT4, B4GalT2, B4GalT3 knockout. The mgat cluster mainly has magt4B, mgat4A/4B, magt5 knockout. The st3gal4/6 knockout forms one cluster. The cluster with joint knockout has KO_st3gal4/6_mgat4a/4b/5, KO_mgat4A/4B/5_B3gnt2, KO_mgat4A/4B/5, KI_st3gal6; KO_mgat4a/4b/5_st3gal4/6. The glycoprofile with Fuc 8, mgat2, mgat 1 forms individual cluster.

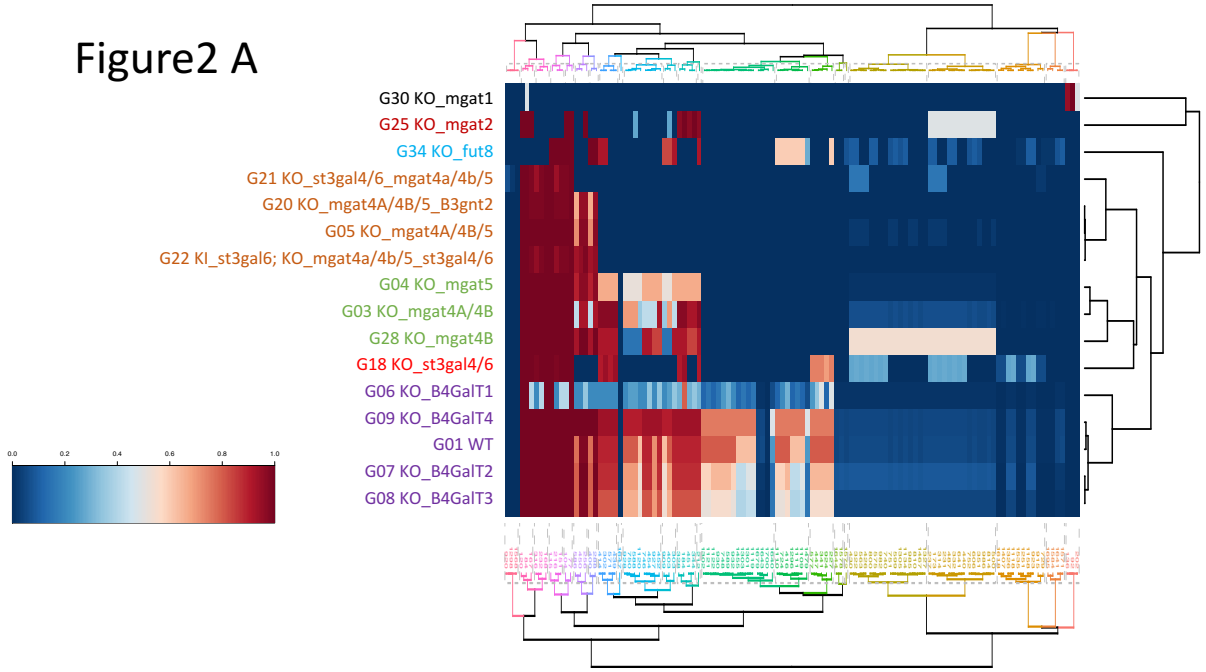
The clustering with annotated glycan is bad because there are only a few glycans.

Figure 2B/Figure2C

- The clustering result is very consistent with NCM paper (Figure2 B). NCM clusters the glycoprofile with the native mass spectrum data. The disagreement comes with the rearrange of the KO mgat2 and KO fu8. Our method take care of the structure difference across isomers. We can distinguish very different structure from very similar molecular mass. Thus, we can find the structure profile of KO mgat2 and KO fut8 are very different from the rest of glycoprofile. Our result is better than NCM ‘s result.

- On the same time, the glycans abundant table are also used for clustering as a comparison. (Figure2 C) It is worse than the clustering with the glyco-substructure. The glycan abundant table used for clustering is very sparse. The cluster is mainly based on the existence of the glycans and the distances between profiles are not well linearly characterized. The cluster is not consistent with the NCM and our result.

Figure2 A



Left side is the glycoprofiles in the NCM paer, the right side is the shared glycoprofile from our pool. The major difference comes from the position of the KOfu8.

77 glycan structure

Figure2 B

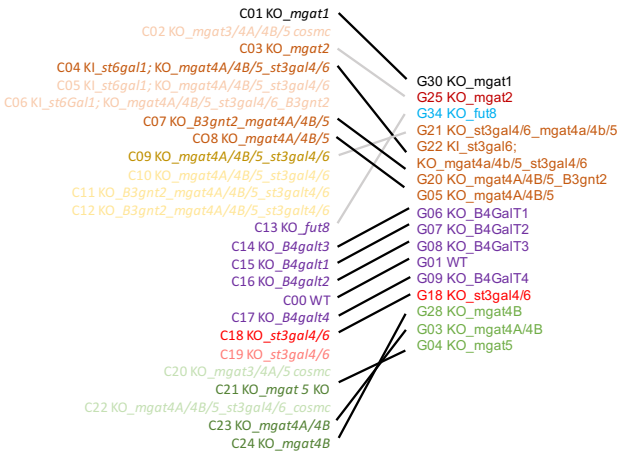
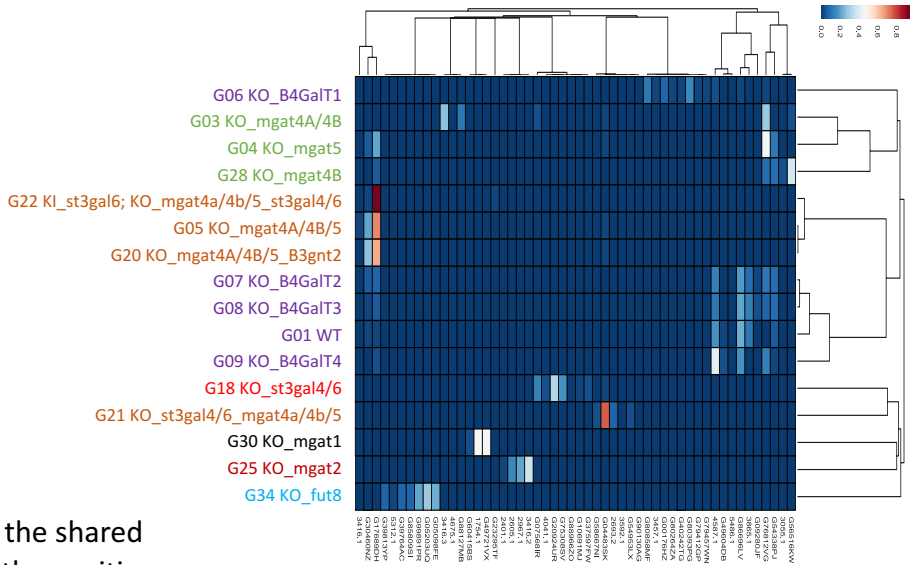


Figure2 C



Left side is the glycoprofiles in the NCM paer, the right side is the shared glycoprofile from our pool. The major difference comes from the position of the KOfu8.

Result 1.2 Characterization of substructure abundance change

- By setting the threshold for cutting the motif cluster. 24 clusters of substructure are picked and the representative substructure for each cluster are generated(see supplementary). The mean abundance of a cluster is used to represent the abundance of the representative substructure. This is the first time to have a comprehensive, automatic quantification of the abundance variation of substructure across multiple glycoprofile.

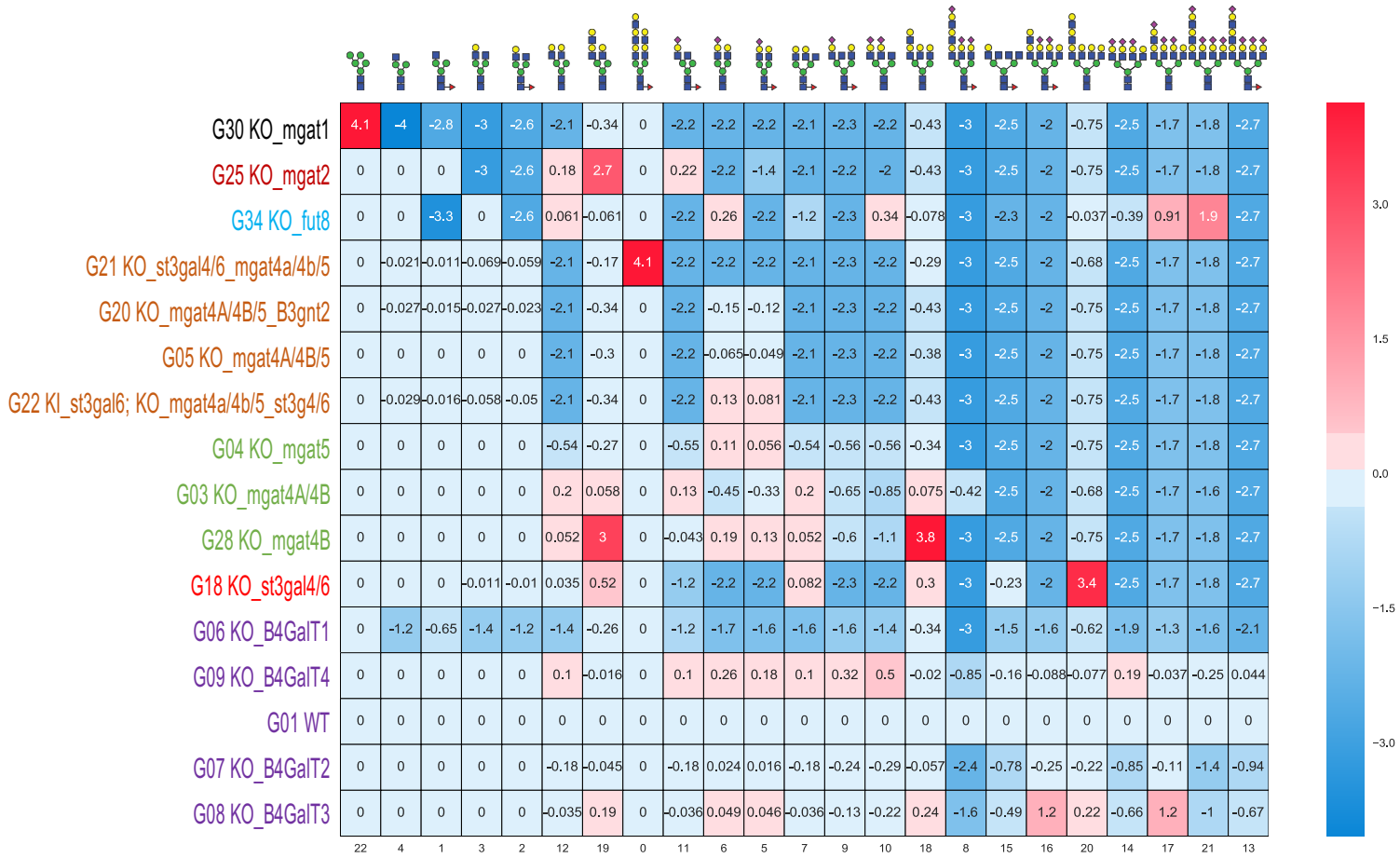


Figure2 D: We are able to quantify the relative abundance of the representative substructure across 16 glycoprofiles. Each column is rescaled with z score. And adjusted with the WT abundance(the abundance of the WT is 0). The substructures from left to right are reordered from single-antennary to tetra-antennary. It shows the decreasing of the complex glycan structure when the knockout becomes more complicate, from bottom to the top. The substructure in deep red implies the significantly increase and several of them have not been noticed before.

Figure 2 EFG

- Figure2 E shows in the KO_fu8 profile, the relative abundance of structures without fucose have significant decrease. But the tetra-antennary polyLacNac elongated N-glycan with no fucose increases as well as the tri-antennary one.
- Figure2 F shows in the KO_stgal4/6 profile, the relative abundance of structures with sialyation have significant decrease. But the tetra-antennary and tri-antennary polyLacNac elongated N-glycan with no sialyation increase.
- Figure2 G shows the knockout in mgat 4B, mgat 4A/4B and mgat 5. Most of the tetra-antennary structures decrease. The Mgat 4B and mgat 5 both have significantly decrease in tri-antennary polyLacNac elongated N-glycan. While mgat 4B has significantly increase in tri-antennary LacNac elongated N-glycan.

Figure2 E

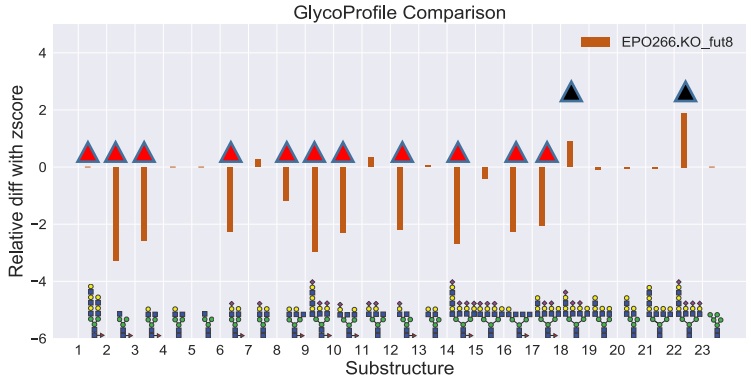


Figure2 F

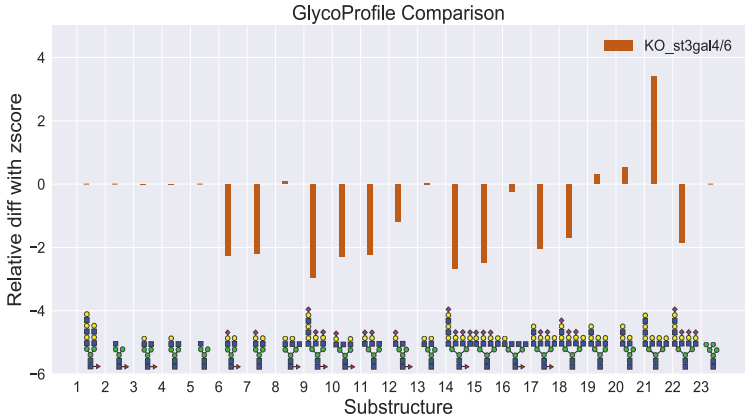
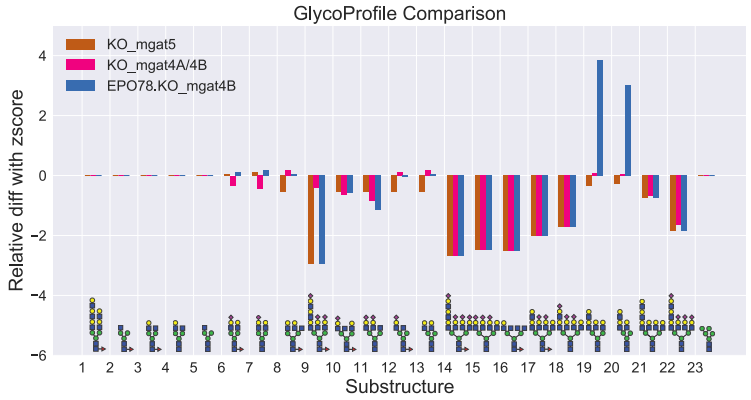
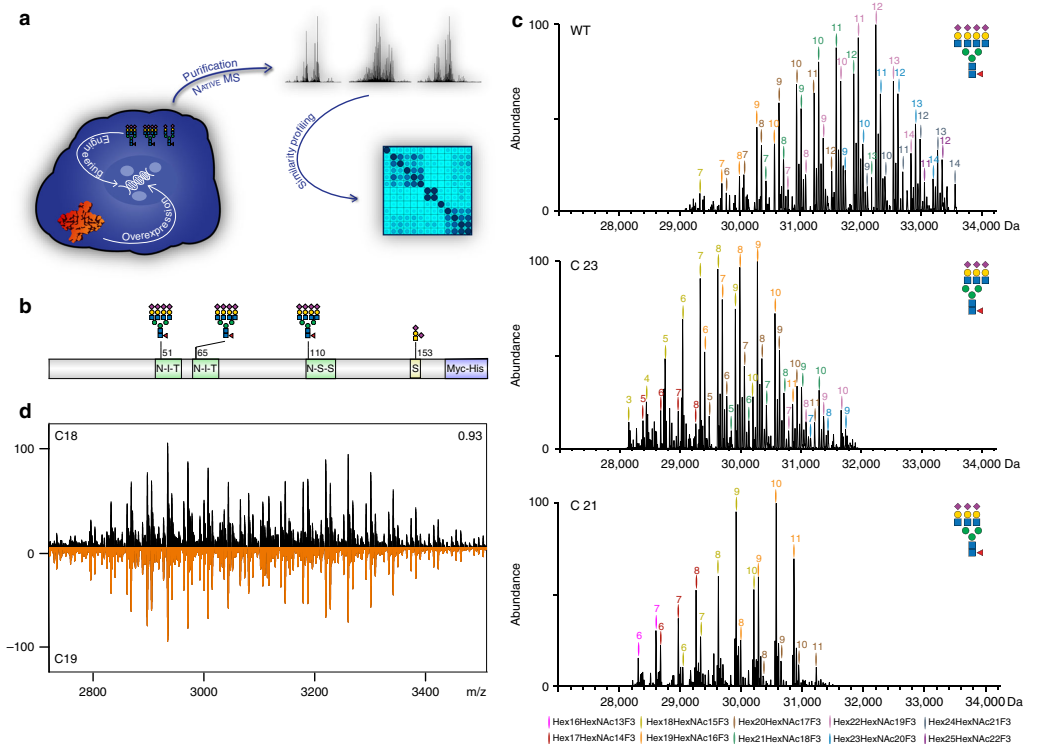


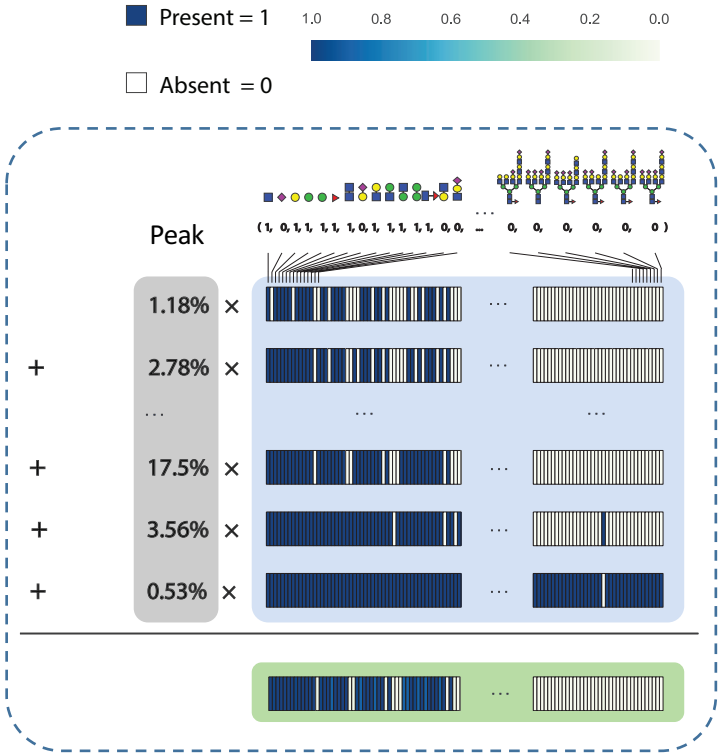
Figure2 G



Data 1 Part 3 Inferring the glycoprofile abundance vector from native mass spectrum (ongoing)

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-05536-3 ARTICLE



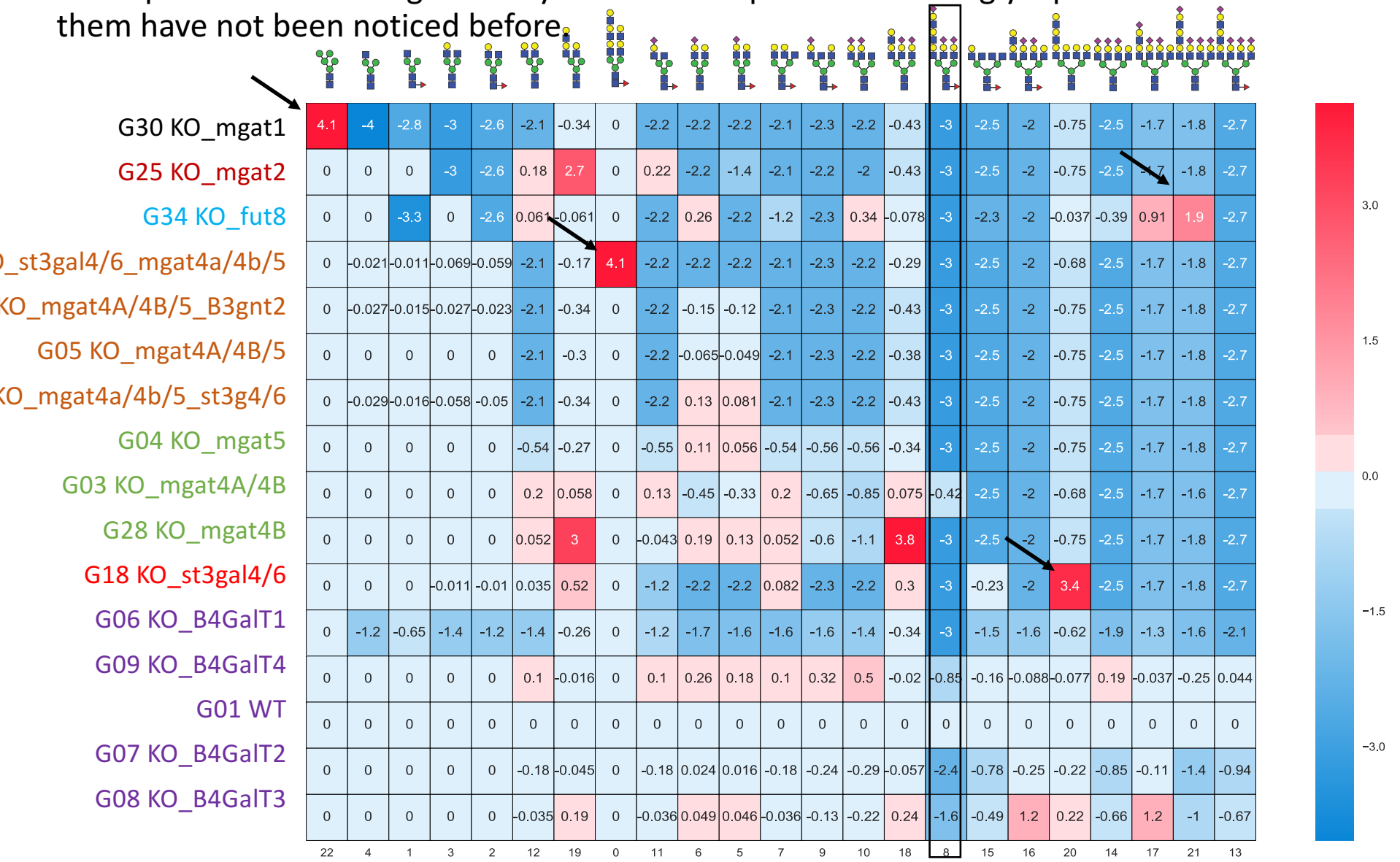


The plot on the left side comes from the **Nature Communication** paper, which shows the native mass spectrum profile for WT, KO mgat4A, KO mgat4B profile. The plot on the right shows it can be transferred to motif vector.

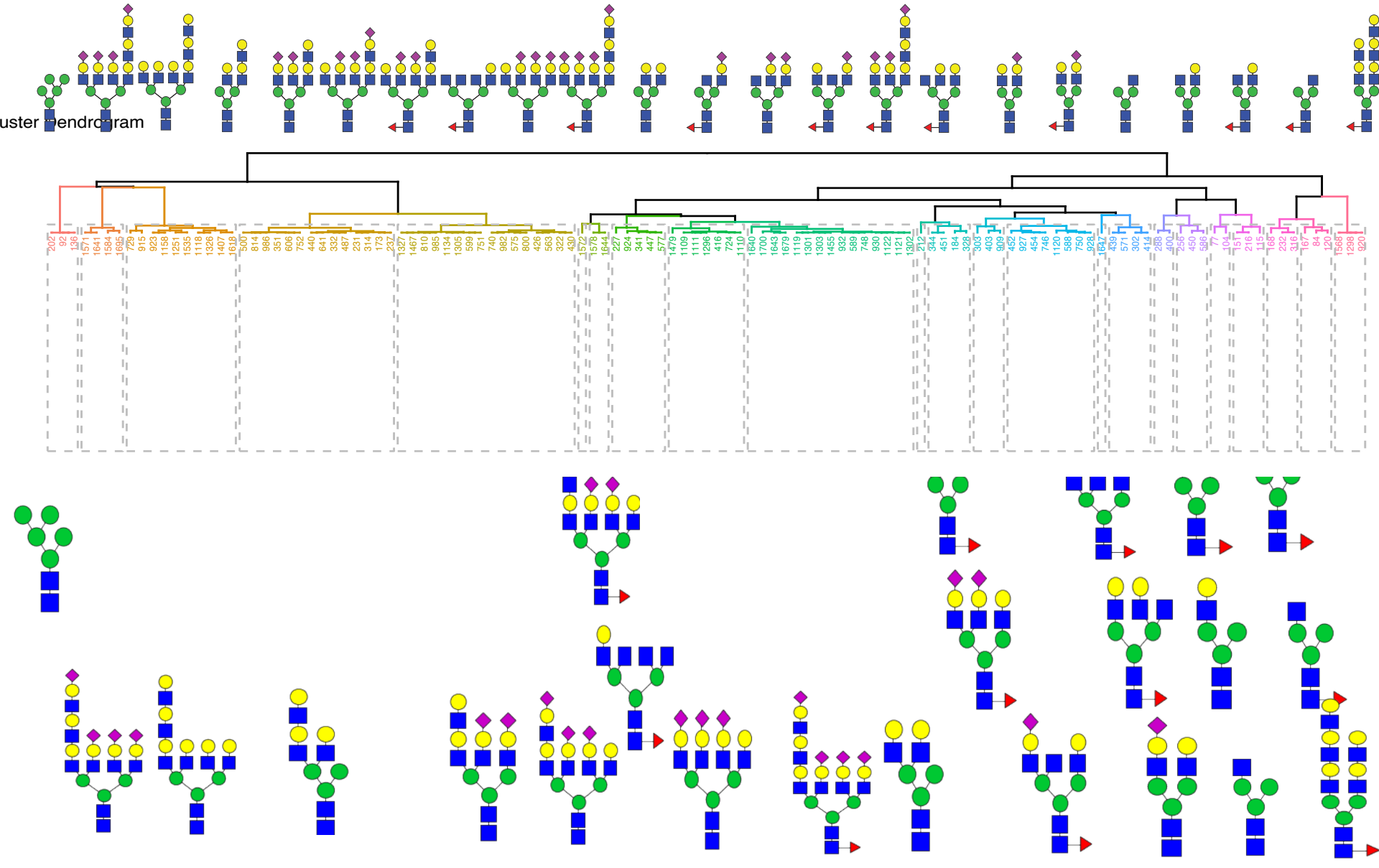
Green box : glycoprofile abundance vector
Grey box: glycan abundance
Blue box: motif structure for each signal peak
This plot shows that how the native mass spectrum data is transformed to glycan **Substructure Vector** and **Glycoprofile Abundance Vector**

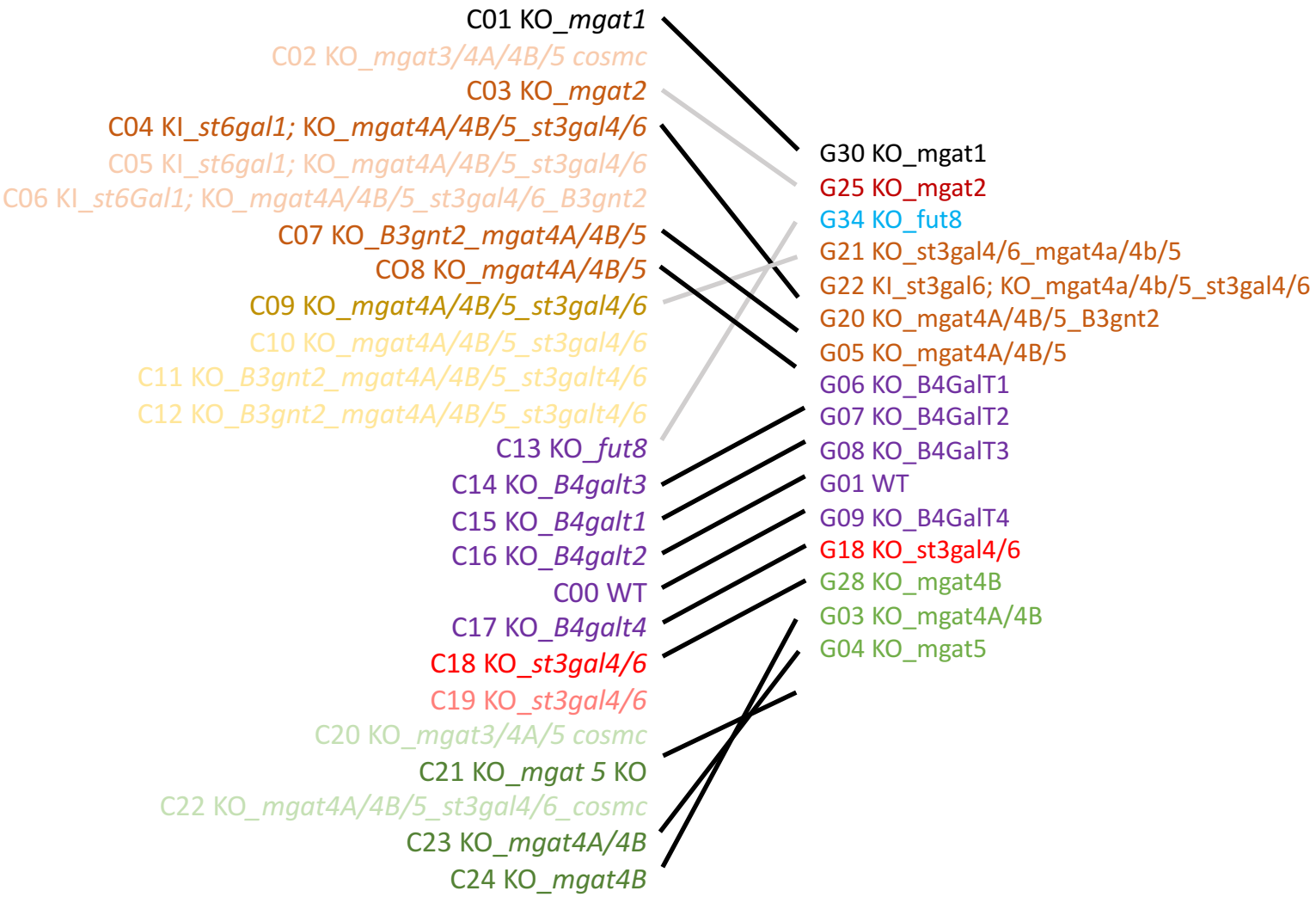
Data 1 part2

Note: each column is rescaled with z_score. And adjusted with the WT abundance. The substructures from left to right are becoming more complicate. It shows the decreasing of the complex glycan structure when the knockout becomes more complicate. Several substructure in deep red means the significantly increase compare with other glycoprofiles and several of them have not been noticed before.

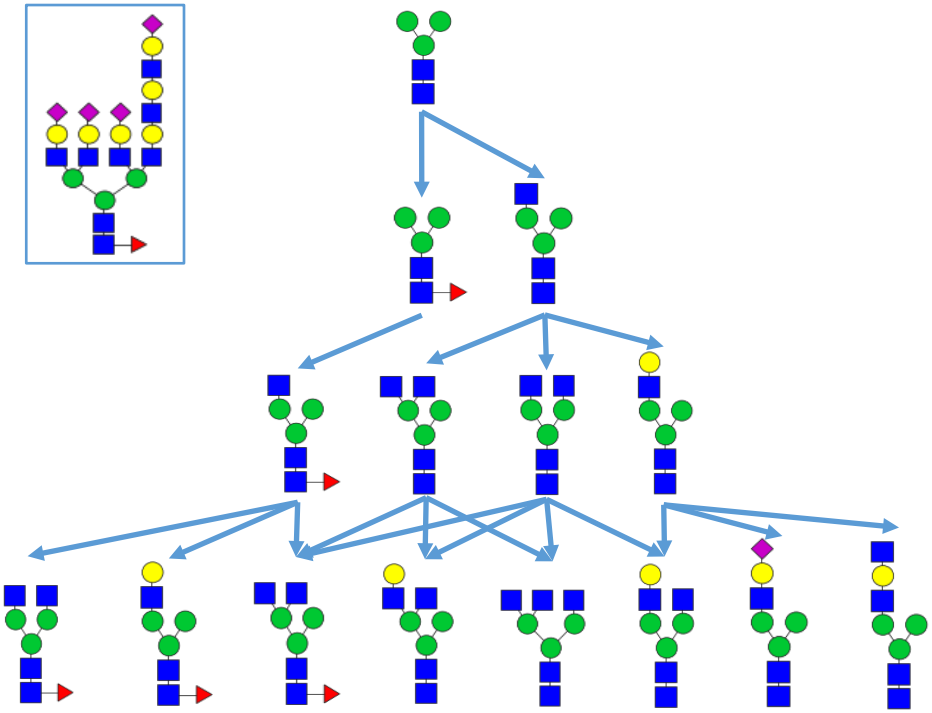


(Supplementary Figure)

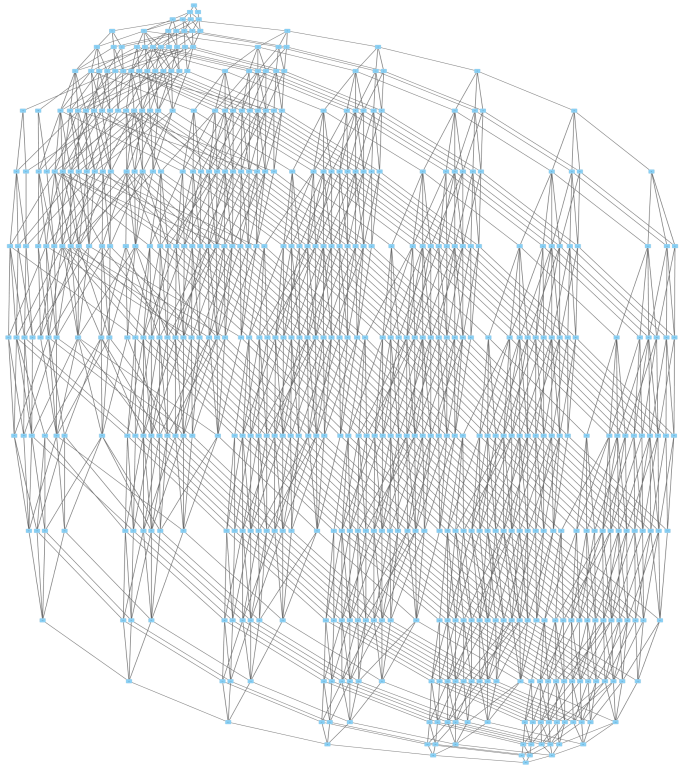



C01 KO_mgat1
C02 KO_mgat3/4A/4B/5 cosmc
C03 KO_mgat2
C04 KI_st6gal1; KO_mgat4A/4B/5_st3gal4/6
C05 KI_st6gal1; KO_mgat4A/4B/5_st3gal4/6
C06 KI_st6Gal1; KO_mgat4A/4B/5_st3gal4/6_B3gnt2
C07 KO_B3gnt2_mgat4A/4B/5
CO8 KO_mgat4A/4B/5
C09 KO_mgat4A/4B/5_st3gal4/6
C10 KO_mgat4A/4B/5_st3gal4/6
C11 KO_B3gnt2_mgat4A/4B/5_st3galt4/6
C12 KO_B3gnt2_mgat4A/4B/5_st3galt4/6
C13 KO_fut8
C14 KO_B4galt3
C15 KO_B4galt1
C16 KO_B4galt2
C00 WT
C17 KO_B4galt4
C18 KO_st3gal4/6
C19 KO_st3gal4/6
C20 KO_mgat3/4A/5 cosmc
C21 KO_mgat 5 KO
C22 KO_mgat4A/4B/5_st3gal4/6_cosmc
C23 KO_mgat4A/4B
C24 KO_mgat4B
G30 KO_mgat1
G25 KO_mgat2
G34 KO_fut8
G21 KO_st3gal4/6_mgat4a/4b/5
G22 KI_st3gal6; KO_mgat4a/4b/5_st3gal4/6
G20 KO_mgat4A/4B/5_B3gnt2
G05 KO_mgat4A/4B/5
G06 KO_B4GalT1
G07 KO_B4GalT2
G08 KO_B4GalT3
G01 WT
G09 KO_B4GalT4
G18 KO_st3gal4/6
G28 KO_mgat4B
G03 KO_mgat4A/4B
G04 KO_mgat5

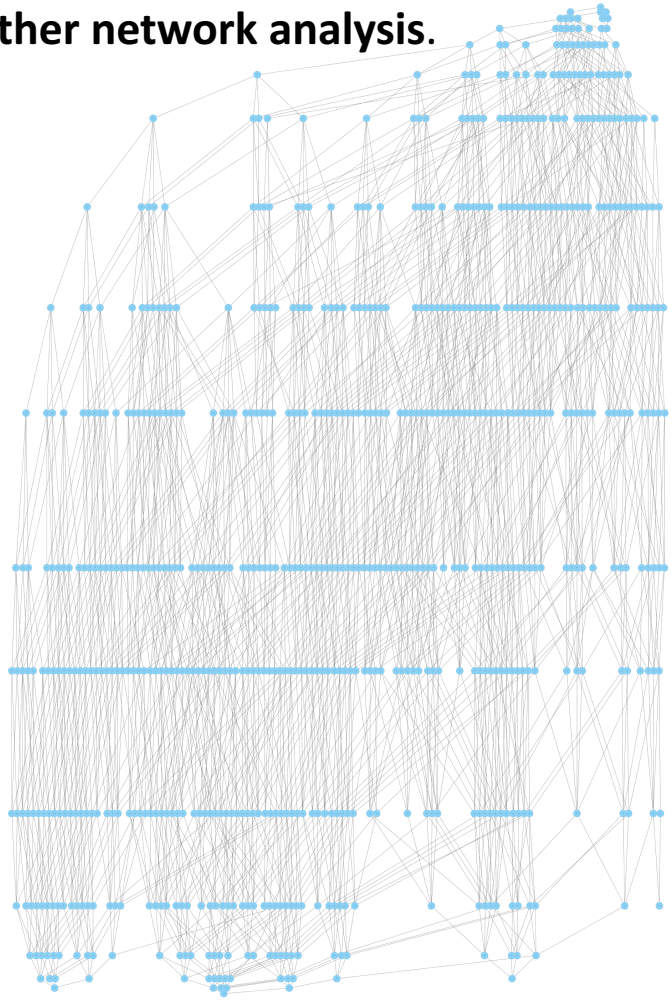
Note: a glycan can be break down to substructure and then form a synthesis network. The left side shows the top several substructures from the right side network.

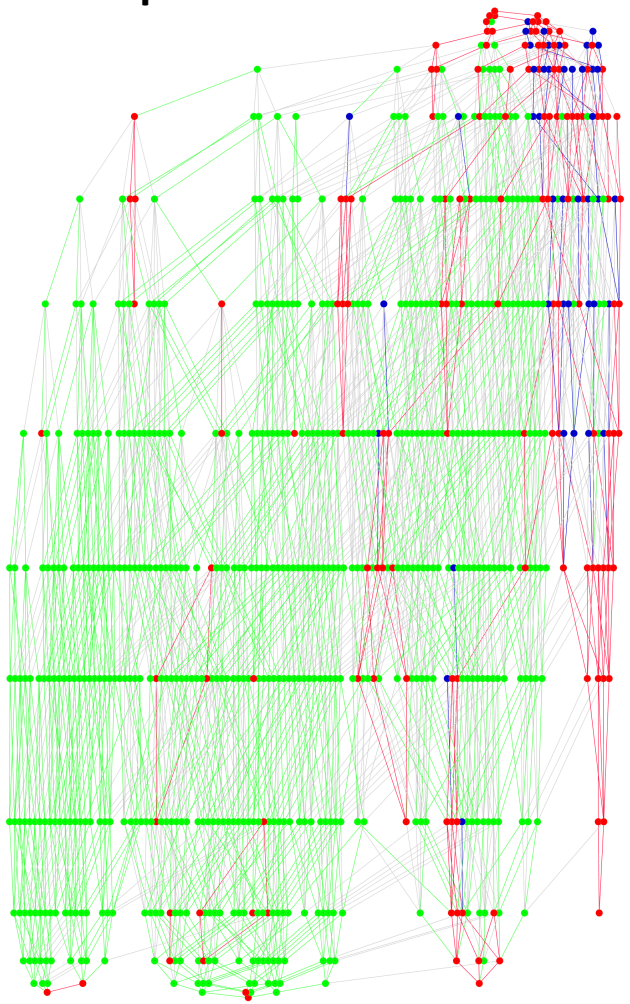


...



Note: multiple glycans' synthesis network can be merged together and multiple glycoprofiles can also be merged together. Thus, the comparison across the glycoprofiles is feasible and the important change of substructure abundance can be highlighted (red and blue nodes). **Only the nodes with blue and red color are picked for further network analysis**.



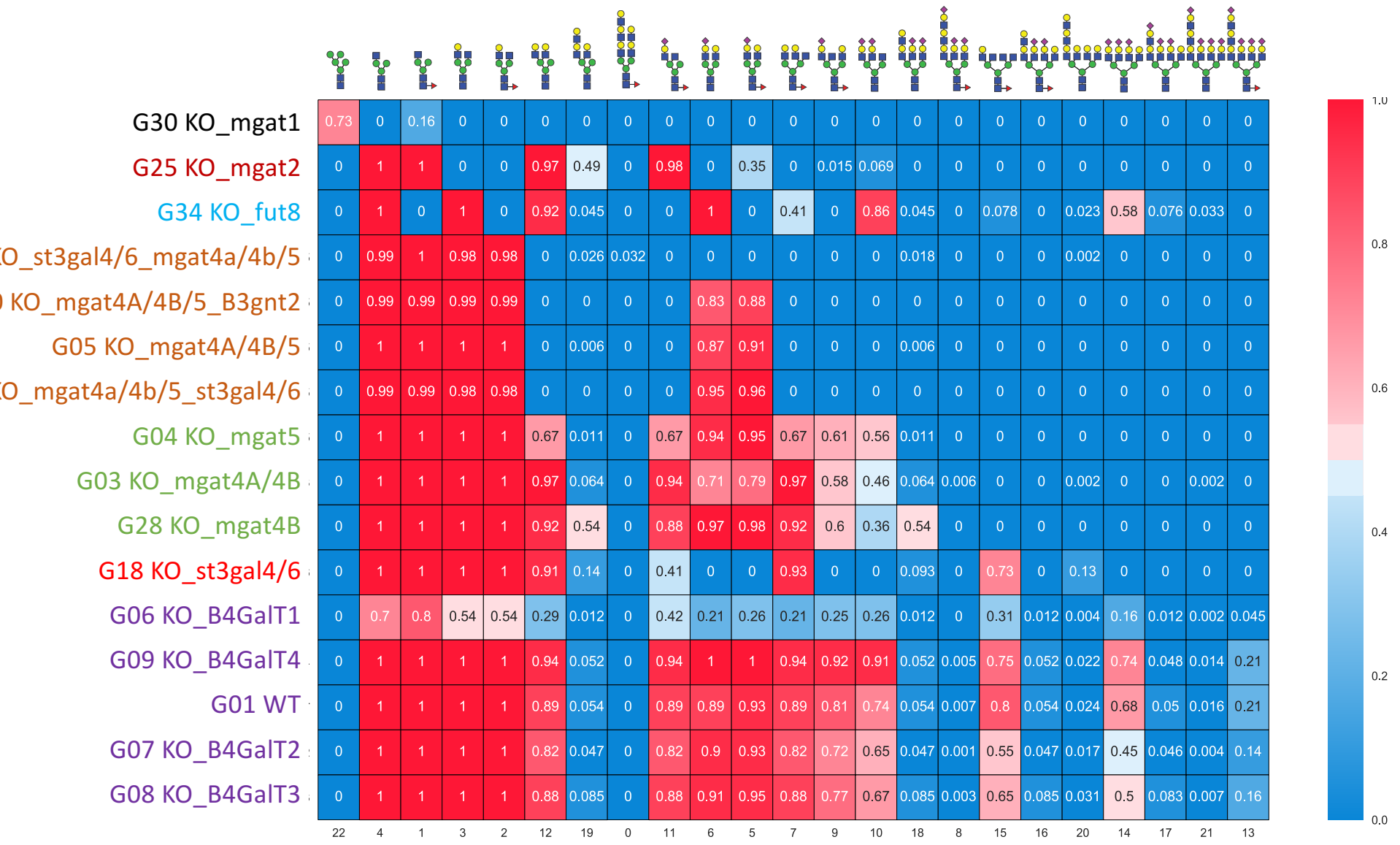


GlyCompare

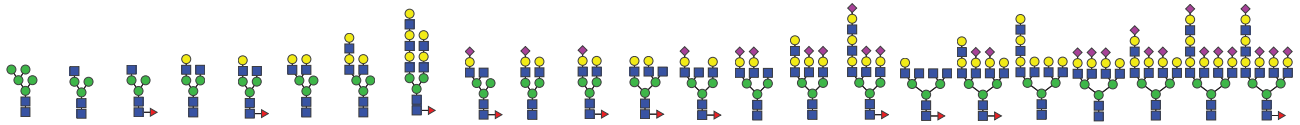
- Comparing large set of glycoprofiles
- Tracking the substructure(motif) abundance change
- Correlating the phenotypes with the motif’s abundance.
- Illustrating the perturbation of the synthesis netowork.

Data 1 part2

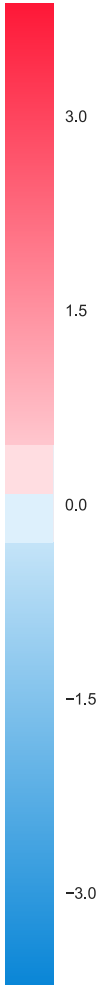
Note: the mean of each cluster is showed here.



Note: each column is rescaled with z_score.

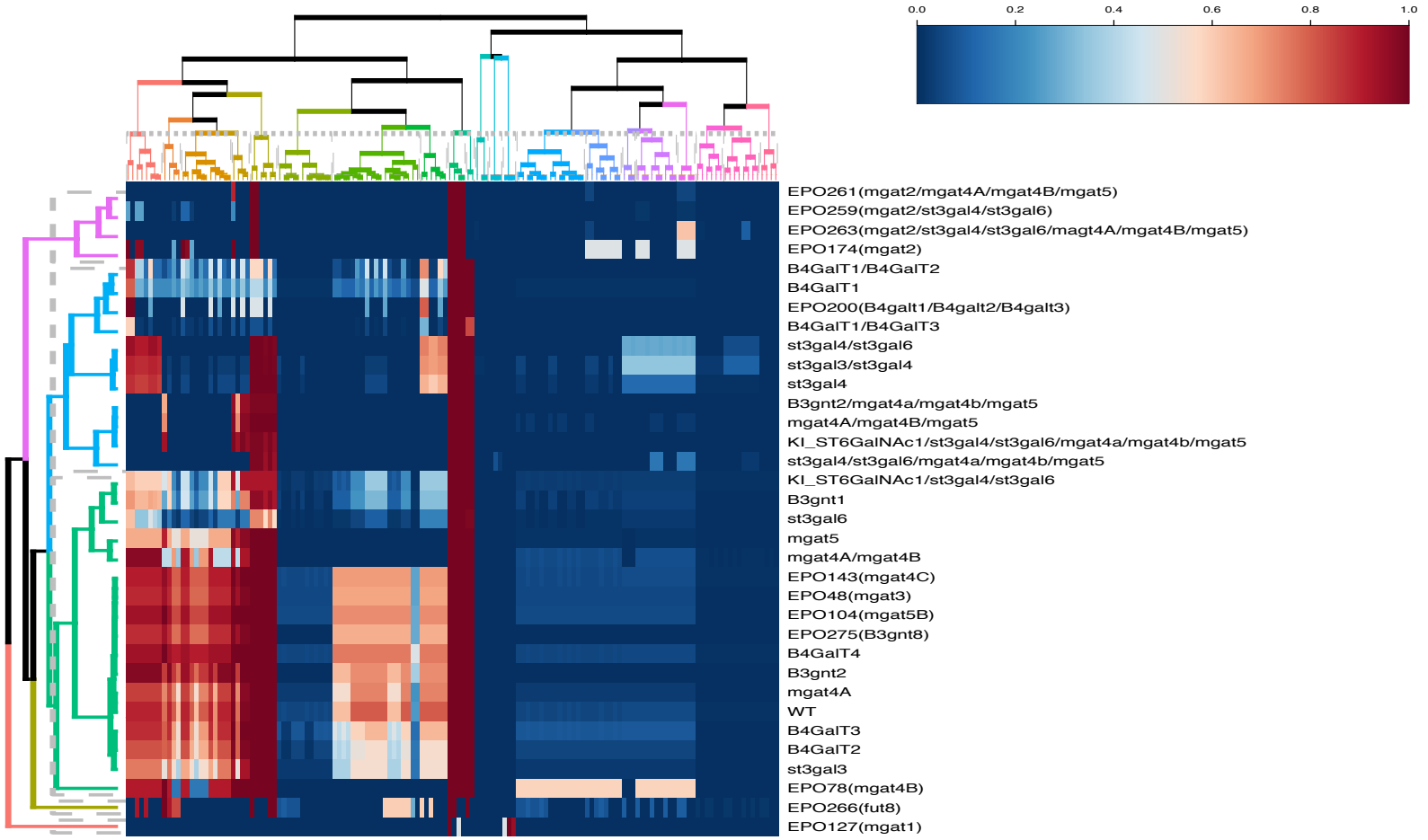


	22	4	1	3	2	12	19	0	11	6	5	7	9	10	18	8	15	16	20	14	17	21	13
G30 KO_mgat1	3.9	-3.7	-2.3	-2.5	-2	-1.4	-0.61	-0.26	-1.2	-1.6	-1.5	-1.1	-0.93	-1	-0.51	-0.59	-0.76	-0.59	-0.5	-0.7	-0.68	-0.56	-0.61
G25 KO_mgat2	-0.26	0.33	0.42	-2.5	-2	0.95	2.4	-0.26	1.2	-1.6	-0.64	-1.1	-0.89	-0.83	-0.51	-0.59	-0.76	-0.59	-0.5	-0.7	-0.68	-0.56	-0.61
G34 KO_fut8	-0.26	0.33	-2.9	0.47	-2	0.83	-0.33	-0.26	-1.2	0.89	-1.5	-0.17	-0.93	1.5	-0.15	-0.59	-0.51	-0.59	0.21	1.4	1.9	3.2	-0.61
O_st3gal4/6_mgat4a/4b/5	-0.26	0.31	0.41	0.4	0.5	-1.4	-0.44	3.9	-1.2	-1.6	-1.5	-1.1	-0.93	-1	-0.37	-0.59	-0.76	-0.59	-0.43	-0.7	-0.68	-0.56	-0.61
KO_mgat4A/4B/5_B3gnt2	-0.26	0.3	0.41	0.44	0.54	-1.4	-0.61	-0.26	-1.2	0.47	0.63	-1.1	-0.93	-1	-0.51	-0.59	-0.76	-0.59	-0.5	-0.7	-0.68	-0.56	-0.61
G05 KO_mgat4A/4B/5	-0.26	0.33	0.42	0.47	0.56	-1.4	-0.57	-0.26	-1.2	0.56	0.7	-1.1	-0.93	-1	-0.46	-0.59	-0.76	-0.59	-0.5	-0.7	-0.68	-0.56	-0.61
O_mgat4a/4b/5_st3gal4/6	-0.26	0.3	0.41	0.41	0.51	-1.4	-0.61	-0.26	-1.2	0.76	0.83	-1.1	-0.93	-1	-0.51	-0.59	-0.76	-0.59	-0.5	-0.7	-0.68	-0.56	-0.61
G04 KO_mgat5	-0.26	0.33	0.42	0.47	0.56	0.23	-0.54	-0.26	0.43	0.74	0.81	0.46	0.79	0.63	-0.42	-0.59	-0.76	-0.59	-0.5	-0.7	-0.68	-0.56	-0.61
G03 KO_mgat4A/4B	-0.26	0.33	0.42	0.47	0.56	0.96	-0.21	-0.26	1.1	0.18	0.42	1.2	0.71	0.34	-0.002	1.9	-0.76	-0.59	-0.44	-0.7	-0.68	-0.35	-0.61
G28 KO_mgat4B	-0.26	0.33	0.42	0.47	0.56	0.82	2.7	-0.26	0.94	0.82	0.88	1.1	0.76	0.052	3.8	-0.59	-0.76	-0.59	-0.5	-0.7	-0.68	-0.56	-0.61
G18 KO_st3gal4/6	-0.26	0.33	0.42	0.46	0.55	0.8	0.25	-0.26	-0.2	-1.6	-1.5	1.1	-0.93	-1	0.22	-0.59	1.5	-0.59	3.7	-0.7	-0.68	-0.56	-0.61
G06 KO_B4GalT1	-0.26	-0.87	-0.23	-0.9	-0.62	-0.67	-0.53	-0.26	-0.17	-1	-0.87	-0.64	-0.22	-0.26	-0.41	-0.59	0.21	-0.14	-0.38	-0.13	-0.27	-0.3	-0.031
G09 KO_B4GalT4	-0.26	0.33	0.42	0.47	0.56	0.87	-0.29	-0.26	1.1	0.89	0.93	1.1	1.7	1.7	-0.097	1.5	1.6	1.4	0.17	2	0.99	1	2.1
G01 WT	-0.26	0.33	0.42	0.47	0.56	0.77	-0.27	-0.26	0.98	0.63	0.75	1	1.4	1.2	-0.077	2.4	1.8	1.4	0.24	1.8	1	1.3	2.1
G07 KO_B4GalT2	-0.26	0.33	0.42	0.47	0.56	0.59	-0.32	-0.26	0.8	0.65	0.77	0.82	1.1	0.9	-0.13	-0.013	0.97	1.2	0.026	0.93	0.91	-0.12	1.1
G08 KO_B4GalT3	-0.26	0.33	0.42	0.47	0.56	0.73	-0.08	-0.26	0.95	0.68	0.8	0.96	1.2	0.97	0.17	0.72	1.3	2.6	0.47	1.1	2.2	0.26	1.4

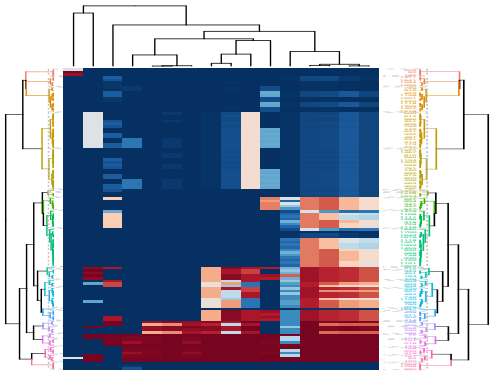


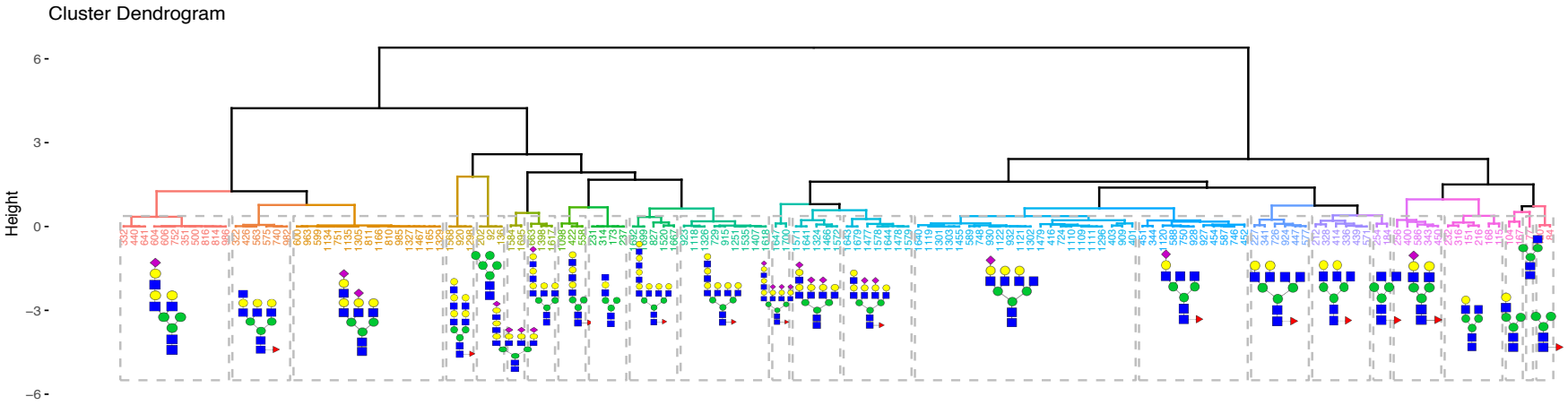
Data 1 part1

Note: preliminary analysis, clustering across 33 glycoprofiles,
based on the abundance of the substructures and glycoprofiles.

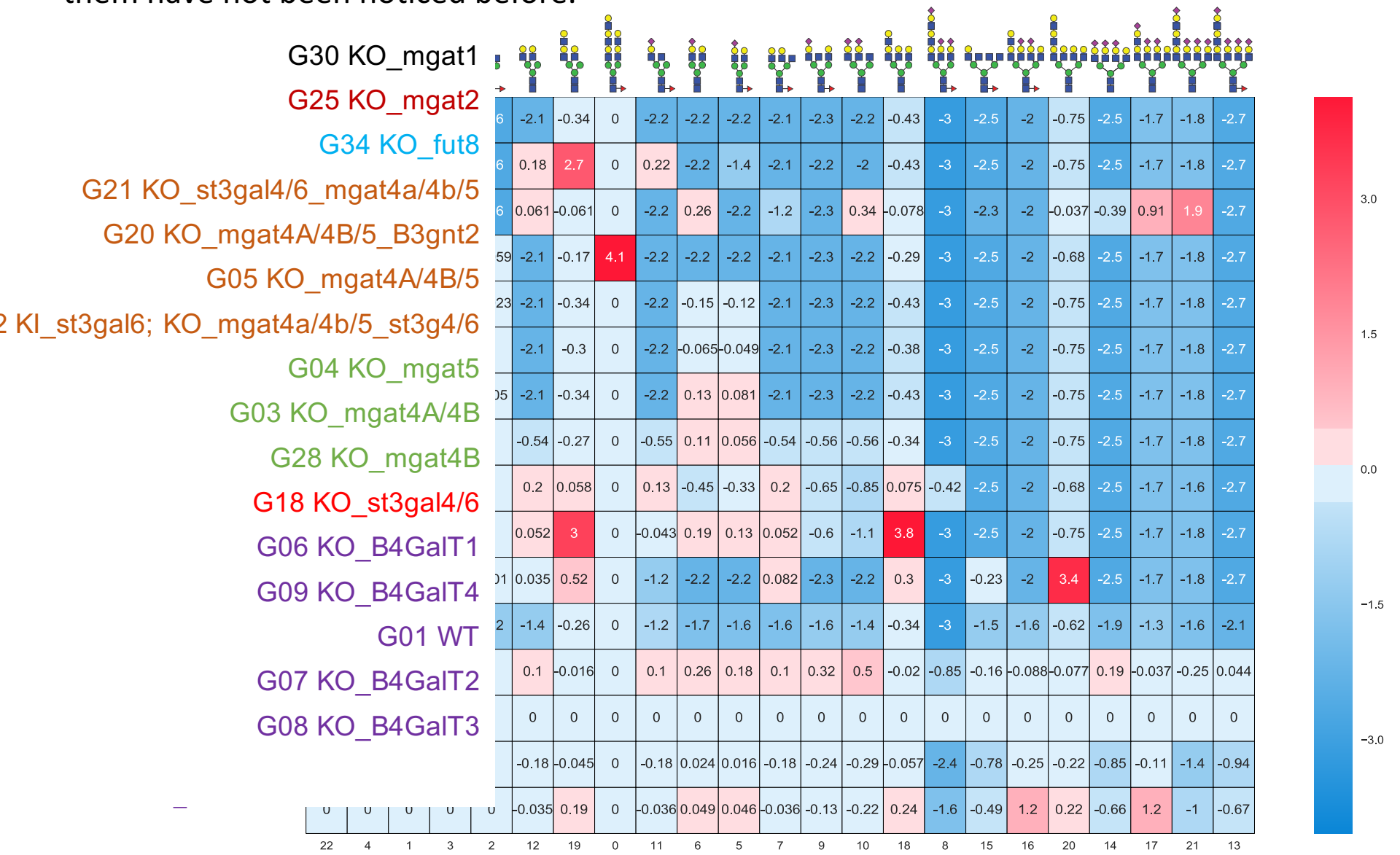


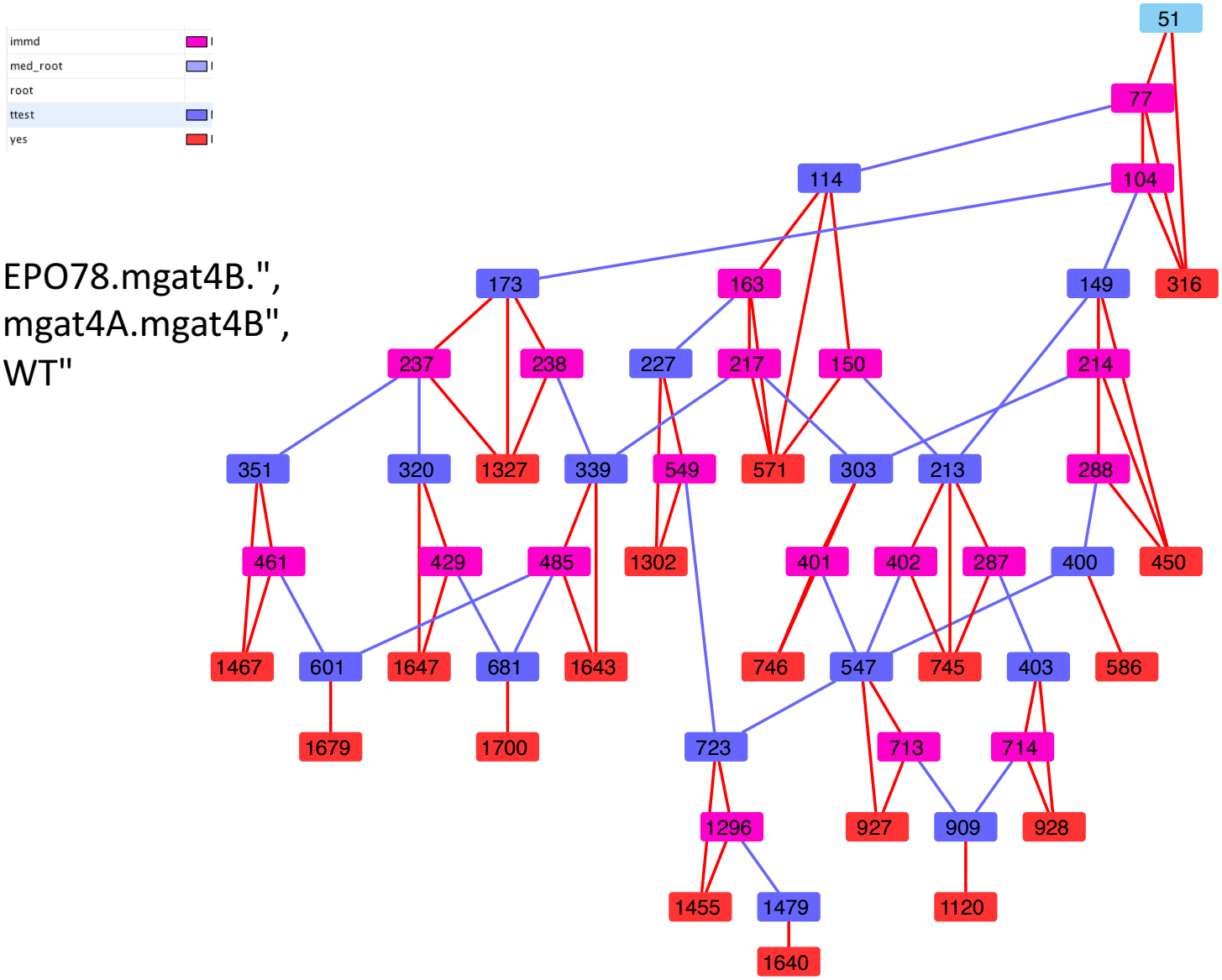
Note: the cluster of the substructure. The representative structures are generated for each cluster.





Note: each column is rescaled with z_score. And adjusted with the WT abundance. The substructures from left to right are becoming more complicate. It shows the decreasing of the complex glycan structure when the knockout becomes more complicate. Several substructure in deep red means the significantly increase compare with other glycoprofiles and several of them have not been noticed before.

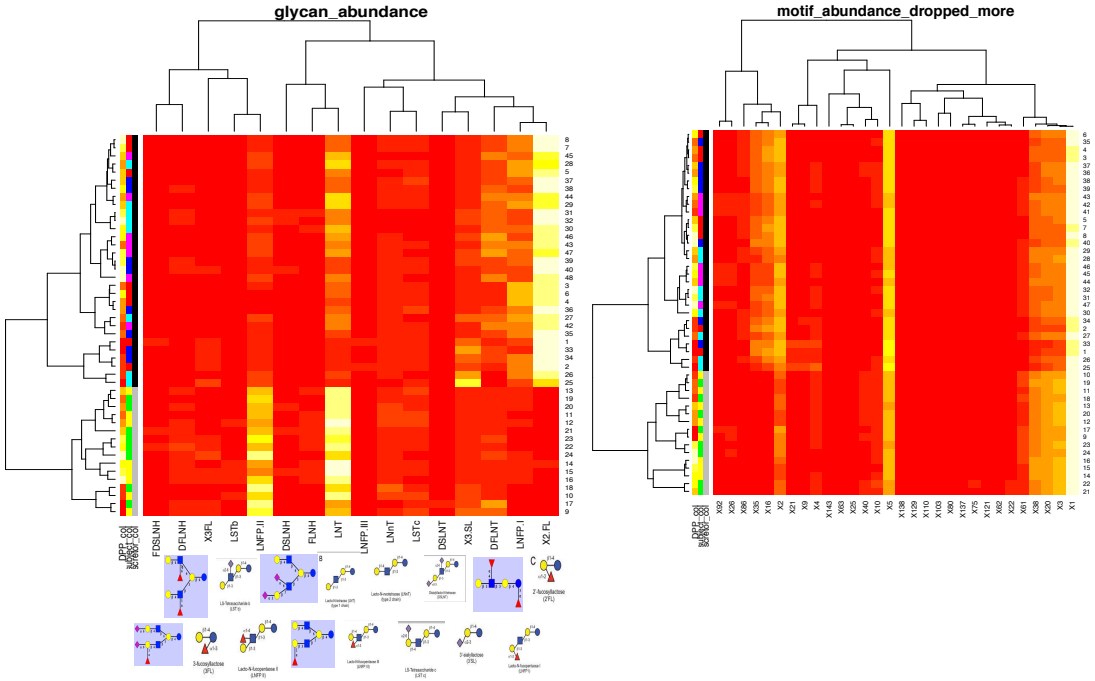


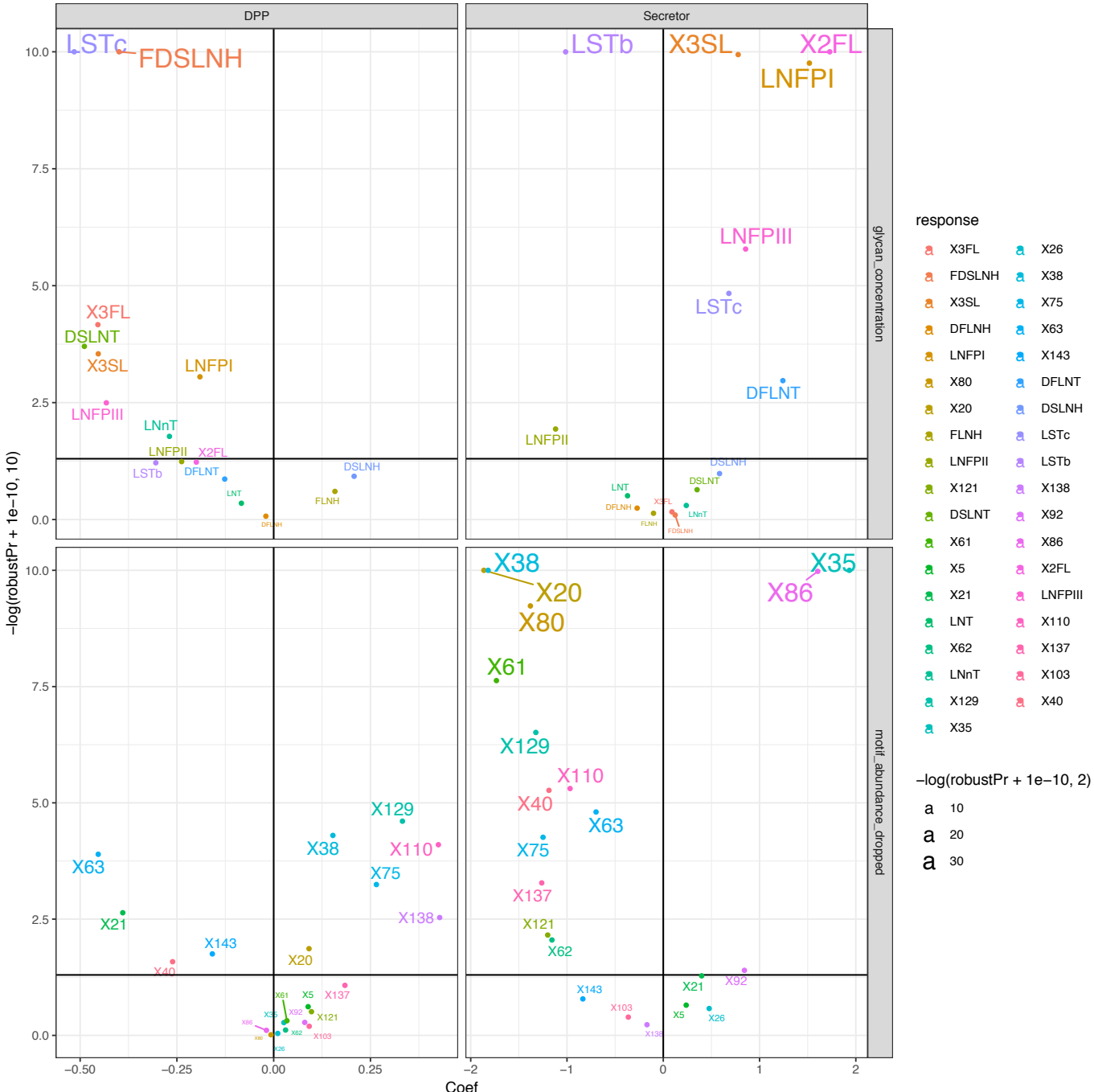

immd
med_root
root
ttest
yes
"EPO78.mgat4B.",
"mgat4A.mgat4B",
"WT"
51
77
114
104
173
163
149
316
237
238
227
217
150
214
351
320
1327
339
549
571
303
213
288
461
429
485
1302
401
402
287
400
450
1467
601
1647
681
1643
746
547
745
403
586
1679
1700
723
713
714
1296
927
909
928
1455
1479
1120
1640

Data 2: HMO data 16 structure, 47 (48-1) glycoprofile

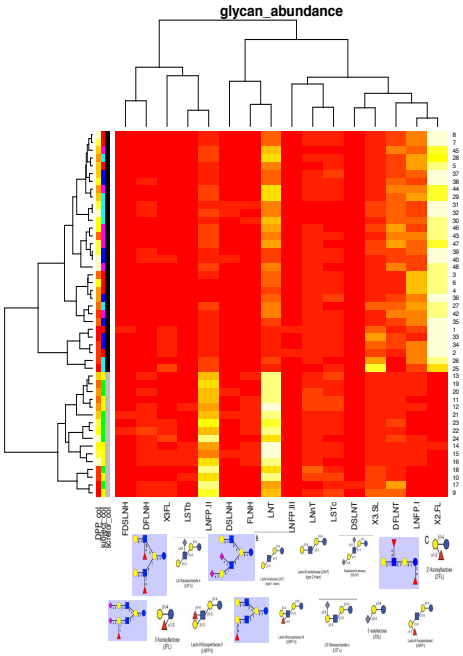
- Results
- Heatmaps
  - Motifs and glycans segregate secretor vs non-secretor
  - Second major organizing trend is time
  - Syalaition decreases over time
  - secretors enriched & non secretors for: 26, 92, 2, 16, 35, 86
- Regression:
  - X80 is anticorrelated with secretor status (log odds, -.015, p=1e-2), X21 is positively correlated with secretor status.
  - X129 & X110 are positively associated with time while syaliated motifs (X?, X??, and X???) are negatively associated with time.
- Scatterplots
  - X80 abundance is anticorrelated with time in non-secretors and positively associated with log time in secretors.
  - This trend is not retrievable in DSLNH, DSLNH and LSTb, the X80 containing HMOs.
- Secretor/non-secretor spearman correlations
  - X?, X??... are positively correlated in secretors and negatively correlated in non-secretors
  - X?, X??... are positively correlated in non-secretors and negatively correlated in secretors
  - motifs positvely correlated in secretors are (fucosylated, fuc & syalaited, galactosylated…?)
  - glycan leve observations
    - LSTb is correlated to LSTc in nonsecretors but not secretors suggesting a secretor-status dependent competition

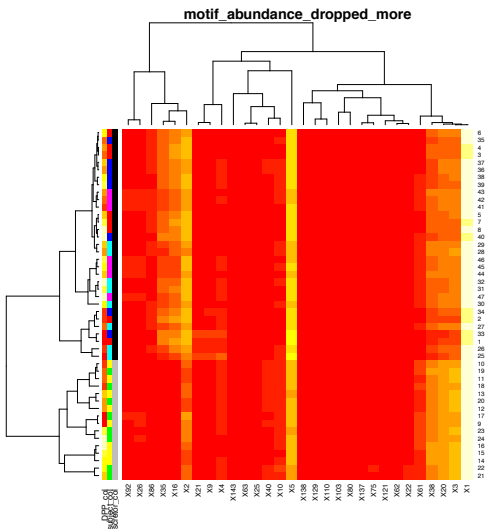
Figure 3 A

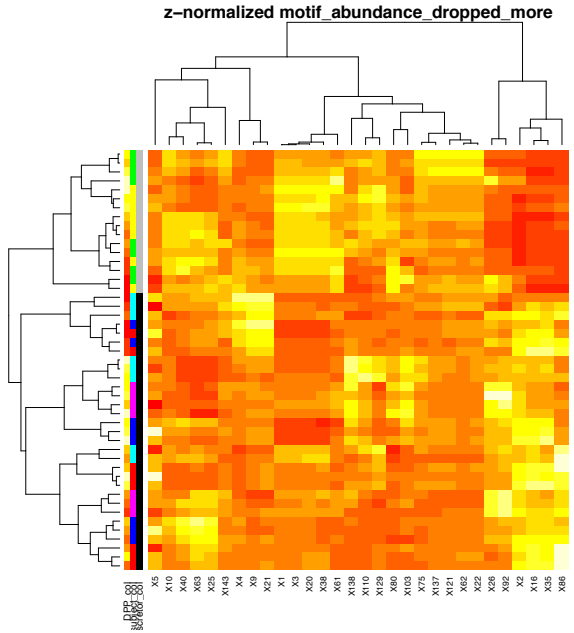

glycan_abundance
motif_abundance_dropped_more
FDSLNH
DFLNH
X3FL
LSTb
LNFP II
DSLNH
FLNH
LNT
LNFP III
LNnT
LSTc
DSLNT
X3.SL
DFLNT
LNFP I
X2.FL

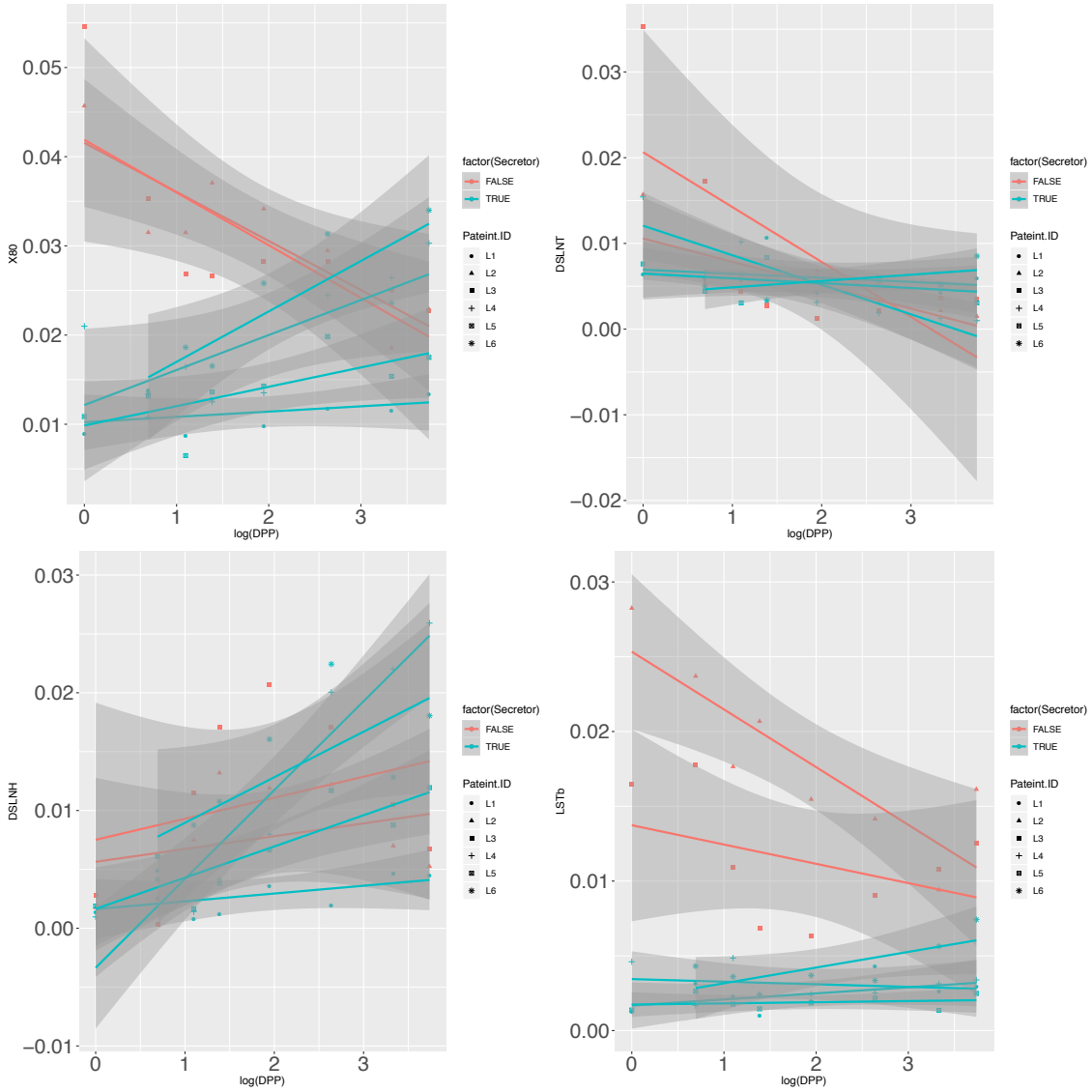

DPP
Secretor
glycan_concentration
motif_abundance_dropped
−log(robustPr + 1e−10, 10)
Coef
response
X3FL
FDSLNH
X3SL
DFLNH
LNFPI
X80
X20
FLNH
LNFPII
X121
DSLNT
X61
X5
X21
LNT
X62
LNnT
X129
X35
X26
X38
X75
X63
X143
X137
X103
X2FL
LNFPIII
LSTc
LSTb
X138
X86
X92
X40
X110
−log(robustPr + 1e−10, 2)
10
20
30

Data 2 part1: Information is more enriched in the motif level. Several motif clusters are highlighted in the plot.



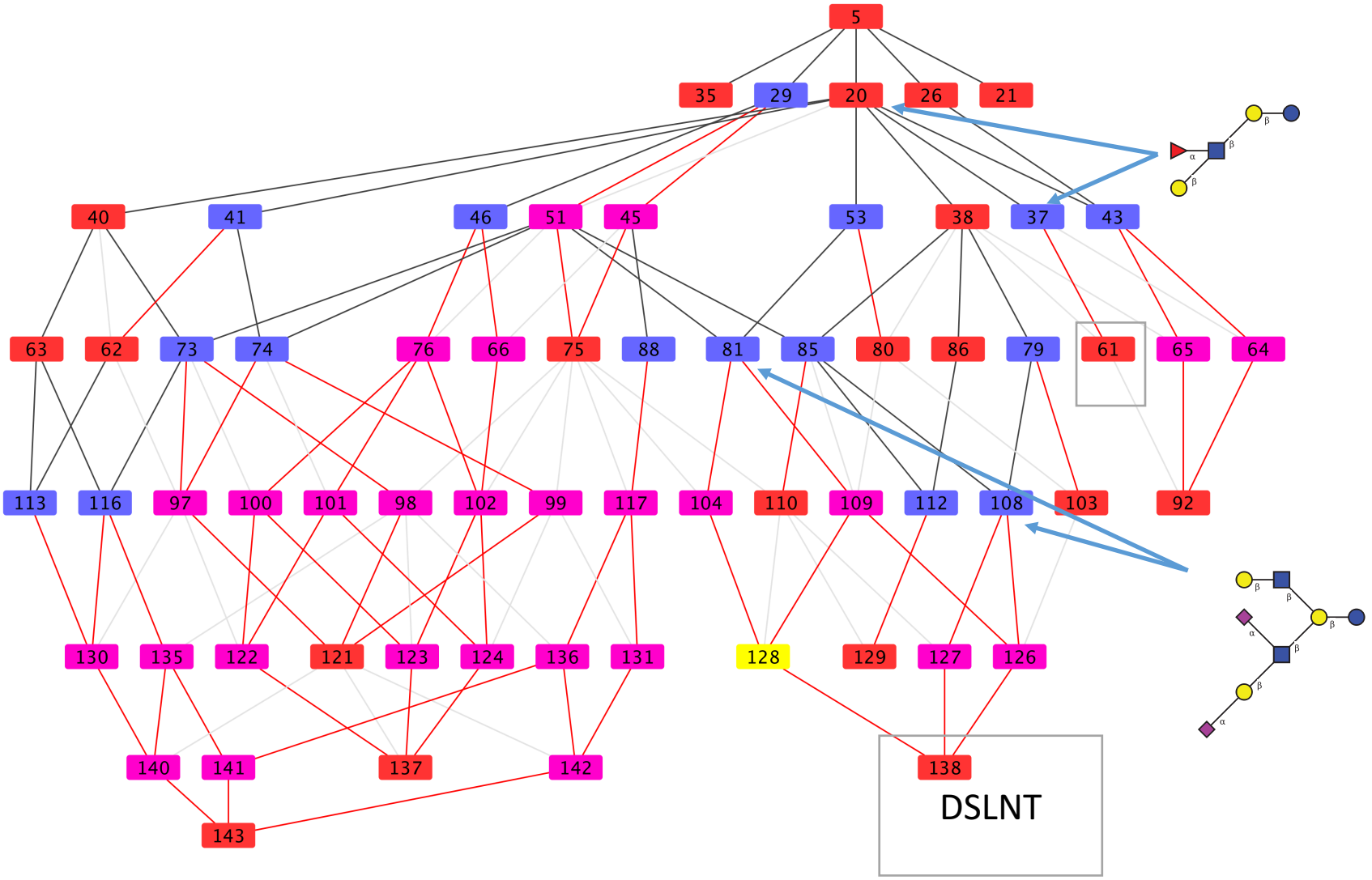





X80
0.05
0.04
0.03
0.02
0.01
0
1
2
3
log(DPP)
DSLNT
0.03
0.02
0.01
0.00
−0.01
−0.02
DSLNH
LSTb
factor(Secretor)
FALSE
TRUE
Pateint.ID
L1
L2
L3
L4
L5
L6

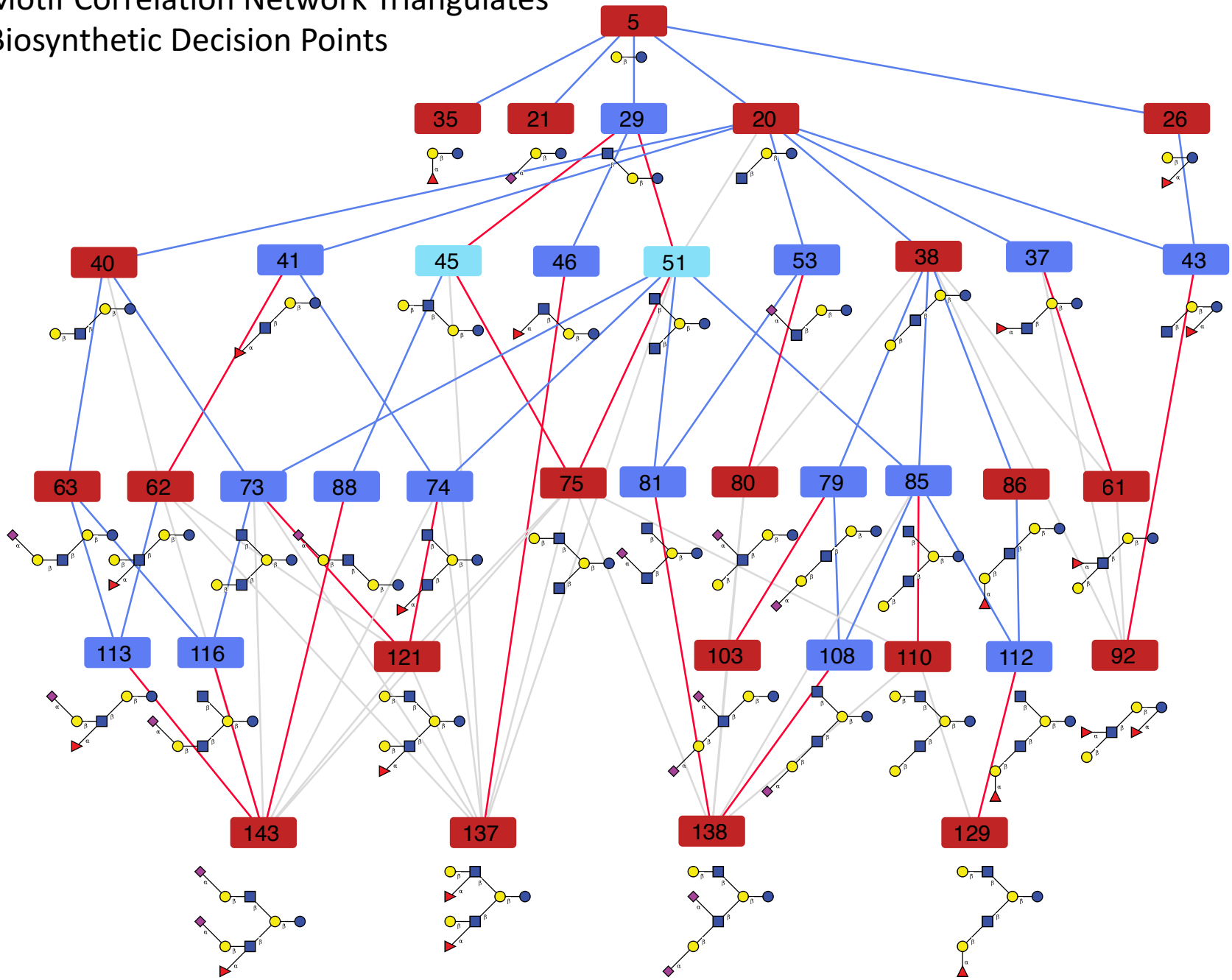
GlyCompare

- Comparing large set of glycoprofiles
- Tracking the substructure(motif) abundance change
- Illustrating the perturbation of the synthesis network.
- Correlating the phenotypes with the motif's abundance.



Motif Correlation Network Triangulates
Biosynthetic Decision Points

Motif Correlation Network Triangulates Biosynthetic Decision Points



GlyCompare

- Comparing large set of glycoprofiles
- Tracking the substructure(motif) abundance change
- Illustrating the perturbation of the synthesis network.
- Correlating the phenotypes with the motif’s abundance.

Data 2 Part 2: Regression model are used to evaluate the importance of motifs as features to predict the Date after birth and Secretor(or not) phenotypes. Motifs that are enriched with DPP, Secretor phenotype. Sialyation is more abundant in the early date after the birth. And x80 motif which is Lst b, have different abundance between the secretor and non secretor.
Y axis is –log(p), x axis is the coefficient value.