**GlyCompare: a python framework to compare the glycoprofile**

Proposal:

Outline:

Abstract:

Introduction:

Result:

Conclusion

Discussion

Supplementary

**1 Introduction**

**1.1 1-2 Glycobiology/Health/importance**

·      Glycan plays a major metabolic, structural and physical roles in biological systems.

·      There are numerous functions the glycan performs such as forming the physical barrier out of cell membrane, changing the water solubility of macromolecules and mediating the protein interaction through specific physical structure properties (Varki 2017).

·      The glycan is a widely used signal molecule outside of the cell membrane and on the surface of the glycoprotein.

·      The change of the glycosylation pattern of tumor will not only change the tumor’s growth and metastasis, but also effect the immune cell activity within the tumor microenvironment. (Rodríguez, Schetters, and Van Kooyk 2018)

·      Change of the glycan structure effect the stability, activity, antigenicity, and pharmacodynamics of the glycoprotein pharmaceuticals in intact organisms (Varki 2017).

**1.2 Motif is important; LewisX, LewisY; Glycan perform its function through motif epitope/N-glycan/O-glycan/Glyco-engineering**

**1 paragraph,**

The glycosylation modulates the function of biologics in a variety of ways. Lack of a Fucose residue on N-glycans in the Fc domain of IgG molecules significantly increase antibody-dependent cell-medicated cytotoxicity. Galactosylation and sialyation play an important role in complement-dependent cytotoxicity and anti-inflammatory activity, respectively. Moreover, sialyation and increased branching of erythropoietin N-glycans increases its serum half-life, while EPO lacking sialyation exhibits neuroprotective role in vivo (Čaval et al. 2018).

Lewis antigens

The glycan structure

·      Motif is a part of glycan structure that repeatedly appeared with recognized function.

·      Sialyl Tn antigen and Sialyl T antigen. LewisX. LewisY

·      Branchness and Fucose(Rodríguez, Schetters, and Van Kooyk 2018).

**1.3 The current stage of the structure characterization technology/ starting point high-throughput analysis of glycoprofile**

1 paragraph

The high throughput of glycoprofiling data becomes accessible, thanks to the advance of the mass spectrum technology. Matrix-assisted laser desorption/ionization (MALDI) time of flight (TOF) mass spectrometry (MS) can rapidly profiles with information on glycan composition

(Holst et al. 2016; Reiding et al. 2014)with high volume and speed. PGC-LC-ESI-MS/MS and MALDI-MSI were used as complementary technique to generate the tissue-specific glycan and are able to profiling the glycan structure distribution in tissue, which can be used in study the cancer development and cell delineation(Hood, Conrads, and Veenstra 2006). Additionally, high resolution native mass spectrometry, Orbitrap EMR, can be used to study the heterogeneity of N-glycosylation on glycoprotein therapeutics. (Čaval et al. 2018).

The high throughput glycoprofile data can be used to compare the products between a large amount of engineered cell lines result for a batch to batch controls, fast identification of the glycan structure and the fast characterization of the glycoprofile is necessary. There are two pioneer methods are proposed to annotated the glycan structure fingerprint and glycoprofile fingerprint. The native mass spectrum is the first data type used to compare the similarity across glycoprofiles (Čaval et al. 2018). But there is not structure/motif information generated. There is no tool automatically compare the motif’s existence and abundance across glycoprofile. Our method is able to characterize the aberrant structure and tracing the change of the glycan substructures’ distribution.

description of this stochastic system. If we break down all glycan into substructures, we can study the progress of this complex network comprehensively.

**1.4 Our method**

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

**1 paragraph**

With our method, we are able to quantify the substructures as the specific fingerprint for each glycoprofiles and use it as a linear arithmetic to compare the similarity across glycoprofiles.

The

We use the relative abundance of substructure to represent the glycoprofile which makes the most use of the synthesis property.

Description of the Concept

·      When we break down the linkage to substructure (with the help of glypy), we are able to characterize the property of each substructure.

·      With all potential substructure given, we are able to transform the native mass spectrum data to substructure abundance profile. It means we are able to compare the glycoprofile on the substructure’s topology level when the native mass spectrum data is given. It saves a lot of time on annotating the glycoprofile on glycan level. (potential)

·      By breaking down the final secreted glycan profile into substructure glycan profile and mapping the substructure back to the glycan synthesize network, we are able to quantitatively characterize the efficacy of the enzyme on each reaction.  We are able to compare the enzyme’s activity between each glycoprofile.

**3 Result**

Glycoprofile deconvolution and cross-profile comparison

The glycans in a glycoprofile were annotated and their abundance were calculated. The glycans were destructed and represented by a ‘substructure vector’ (Figure 1a). Then, a glycoprofile consisting of multiple glycans was transformed to ‘glycoprofile vector’ by summing a set of weighted motif vector (Firgure 1b). Then, a substructure network is built to remove the nodes that will not contain more information (Figure 1c and 1d) and the nodes that contain the most valuable information are selected for glycoprofiles’ classification.

Classification of 16 EPO glycoprofiles

A comprehensive knockout screen of glycosyltransferases genes controlling N-glycosylation in CHO cells has been performed(Yang et al. 2015). In this study, 37 glycoprofiles of N-glycosylation of human erythropoietin (EPO, having three N-glycans with heterogeneous tetra-antennary structures, low poly-LacNAc and α2,3-linked sialic acid capping) was characterized by MALDI-TOF (matrix-assisted laser desorption/ionization–time of flight) profiling of released permethylated N-glycans using literature data to interpret and annotate structures. Then, on the following study, the high-resolution native mass spectrum glycoproteoform profile is used to classify the similarities on 23 glycosylation profiles about EPO (Čaval et al. 2018). Since these two studies shares 16 same knockouts, these 16 glycoprofiles with abundance and structure annotation are curated. The cell lines include the single or joint knockout targeting galactosylation (b4galt1/2/3/4), sialyation capping (st3gal4/6), N-glycan branching (mgat1/2/4A/4B/5) and core \alpha 6-fucosylation(Yang et al. 2015). The classification by Glycompare is compared with the result from native mass spectrum (Čaval et al. 2018) and also compare with the classification by complete glycostructure.

First, a glycoprofile vector with 722 glycan substructure are generated and then reduced to 118 glycan substructures. The cluster map (Figure2 A) shows the glycoprofile are classified based on the complexity of the N-glycan structure preserved. Each cluster has distinguished glycan structure patterns. The mild group that contains WT also has b4galt1/2/3/4/ knockout. The medium group contains mainly magt4b/4a&4a/5 knockout and st3gal4/6 knockout. A severer group that contains joint knockouts of st3gal4/6 and mgat4a/4b/5, knockouts of mgat4A/4B/5 and b3gnt2, knockouts of mgat4A/4B/5 and knock-in of st3gal6, knockouts of mgat4a/4b/5 and st3gal4/6. The severest groups have glycoprofiles with Fuc 8, mgat2, mgat 1 knockout. It means these three glycoprofiles have major different glycol-substructures.

Second, the classification of the glycoprofiles by the glycans are shown. (Figure2 C). Since the glycans with minor variant are considered as two glycans, the structural differences of the glycan in each knockout causes the sparsity of the glycan table which jeopardized the classification. 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.

Third, the classifiation result is consistent with the classification by the glycoproteoform native mass spectrum data. The disagreement comes with the rearrange of the knockouts of mgat2 and the knockouts of fu8. Since our method takes care of the structure difference across isomers. We can distinguish the different structures that has same mass and the variation of structures that effect lots of substructures. Thus, we have a better interpretation of the glycan structure variants across multiple glycoprofile.

Illustrating the changes of structures across glycoprofiles

After setting the distance threshold to classify the substructure clusters, 24 clusters of substructure were classified. Then, the representative substructure (see method, supplement) for each cluster were generated (see supplementary) and their abundance were represented by the mean abundance of the substructure in each cluster (Figure).

Hence, we were able to quantify the relative abundance of the representative substructure across 16 glycoprofiles and highlight the structure that significantly increase or decrease in one profile. In order to highlight the differences, the substructure abundance was rescaled with z score and adjusted with the WT abundance (the abundance of the WT is always 0). The substructures from left to right were reordered from bi-antennary to tetra-antennary. The complexity of glycan structure decreased when the knockout became more complicate, from bottom to the top. The substructures representatives in deep red implied the significant high abundance in one glycoprofile against the rest and several of them have not been quantified before.

In the KO\_mgat1, the high mannose structure is the only structure it has, so we can find the rest of them are zero. In the KO\_mgat2, the structure of bi-antennary on one mannose significantly increased. In the KO\_st3gal4/6\_mgat4a/4b/5, the unique structure, bi-antennary polyLacNac elongated N-glycan pops up.

In the KO\_fuc8 profile (Figure2 E), the relative abundance of structures without fucose have significant decrease. But the tetra-antennary polyLacNac elongated N-glycan with no fucose increases, which is not noticed before.

In the KO\_stgal4/6 profile(Figure2 F), 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.

In the knockout group of mgat 4B, mgat 4A/4B and mgat 5 (Figure2 G), 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.

3.2 HMO data

3.2.1 It is able to compare the HMO with linkage.

3.2.2 synthesis of network?

47 glycoprofile from 6 human mother are collected. For each mother, the milk is collected in 1, 2, 3, 4, 7, 14, 28, 42 days after the date of birth.

The HMO shows that the sialyation is high related with the DPP in the non-secretion and x84 is predictive to help distinguish the non-secretion and secretion with p-value. there is a motif is highly correlated with the

**4 Discussion**

This is the first time to have a comprehensive, automatic quantification of the abundance variation of substructure across multiple glycoprofile.

**2 Data Preprocess**

2.1.1 Loading glycan structrue

All glycan structures in glycoCT formate were either manually curated through GlyTouCan drawing tool or directly downloaded from the GlyTouCan. The glycan with GlycoCT format was then transformed to glypy.glycan object.

2.1.2 Loading glycoprofile

Glycoprofile had mass and peak information. We manually annotated the mass with glycan structure and transformed the peak height to the relative abundance(?).

2.2 Getting motif vector

The substructures of a glycan are exhaustively abstracted, which breaks down each linkage of the glycan and generate all glycan substructure. All substructure abstracted were merged to a with duplicates removed.

2.4 Generating glycoprofile vector

A glycan was matched with motif vector to get the count vector. All glycans in one glycoprofile are weighted with relative abundance and summed together to a glycoprofile vector.

2.5 Building substructure network

It is a directed acyclic graph and each node is a glycan substructure. With a motif vector giving, starting from the root node (root core structure), a child node is supposed to be a glycan that has only one more monosaccharaide adding on its parent node. Thus, one child node can have multiple parent node and vice versa. A child node conditionally depends on the parent nodes since it cannot exist without at least one parent node.

2.6 Reducing the size of glycoprofile vector by substructure network

In a substructure network, the dependences of a child to parent is calculated by the correlation of abundance cross all glycoprofile. If the correlation is 1, the child node depends on the parent node and the parent node’s information is same as child node’s. Thus, the parent node can be removed in the glycoprofile vector when we are doing classification.

2.7 Classifiying the glycoprofile

The glycoprofile vector and the standardized glycoprofile vector can be used for the classification respectively. The pearson correlation with complete distance were used to cluster the glycoprofile. Not only the glycoprofile will be classified, but the substructures. The substructures that have same abundance across multiple glycoprofile will be clustered together.

2.8 Generating the representative substructure

The glycans in one cluster were aligned together and the structure shared by a certain percent, generally 51%, of glycans named as the representation substructure.

* 1. Regressions Model: phenotype ~ substructure

Regression models, specifically Generalized Estimating Equations (GEE), were constructed to examine associations between secretor status, days post-partum (DPP) and motif abundance. For compatibility between motifs, linear models were constructed to predict the z-score normalized motif abundance from an additive model of log(DPP) and secretor status. Subject specific effects are accounted for with the GEE exchangeable correlation structure.

We are abstracting the feature for a

**Compare Result:**

From the structure perspective, we could compare the glycoprofile on both topology and exact structure level.

From the quantitative perspective, we could compare the glycoprofile on the substructure existence level or abundance level.

After the glycan structure is given

Čaval, Tomislav, Weihua Tian, Zhang Yang, Henrik Clausen, and Albert J. R. Heck. 2018. “Direct Quality Control of Glycoengineered Erythropoietin Variants.” *Nature Communications* 9 (1): 3342. https://doi.org/10.1038/s41467-018-05536-3.

Holst, Stephanie, Anna J. M. Deuss, Gabi W. van Pelt, Sandra J. van Vliet, Juan J. Garcia-Vallejo, Carolien A. M. Koeleman, André M. Deelder, et al. 2016. “N-Glycosylation Profiling of Colorectal Cancer Cell Lines Reveals Association of Fucosylation with Differentiation and Caudal Type Homebox 1 (CDX1)/Villin MRNA Expression.” *Molecular & Cellular Proteomics* 15 (1): 124–40. https://doi.org/10.1074/mcp.M115.051235.

Hood, Brian L., Thomas P. Conrads, and Timothy D. Veenstra. 2006. “Mass Spectrometric Analysis of Formalin-Fixed Paraffin-Embedded Tissue: Unlocking the Proteome Within.” *Proteomics* 6 (14): 4106–14. https://doi.org/10.1002/pmic.200600016.

Reiding, Karli R., Dennis Blank, Dennis M. Kuijper, André M. Deelder, and Manfred Wuhrer. 2014. “High-Throughput Profiling of Protein N-Glycosylation by MALDI-TOF-MS Employing Linkage-Specific Sialic Acid Esterification.” *Analytical Chemistry* 86 (12): 5784–93. https://doi.org/10.1021/ac500335t.

Rodríguez, Ernesto, Sjoerd T.T. Schetters, and Yvette Van Kooyk. 2018. “The Tumour Glyco-Code as a Novel Immune Checkpoint for Immunotherapy.” *Nature Reviews Immunology* 18 (3): 204–11. https://doi.org/10.1038/nri.2018.3.

Varki, Ajit. 2017. “Biological Roles of Glycans.” *Glycobiology* 27 (1): 3–49. https://doi.org/10.1093/glycob/cww086.

Yang, Zhang, Shengjun Wang, Adnan Halim, Morten Alder Schulz, Morten Frodin, Shamim H Rahman, Malene B Vester-Christensen, et al. 2015. “Engineered CHO Cells for Production of Diverse, Homogeneous Glycoproteins.” *Nature Biotechnology* 33 (8): 842–44. https://doi.org/10.1038/nbt.3280.