Supporting Information 1

Combining Glucose Units, *m/z* and Collision Cross Section values: Multi-attribute data for increased accuracy in automated glycosphingolipid glycan identifications and its application in Triple Negative Breast Cancer

Katherine Wongtrakul-Kish[1*], Ian Walsh[1], Lyn Chiin Sim[1], Amelia Mak[1], Brian Liau[1], Vanessa Ding[2], Noor Hayati[1], Han Wang[3], Andre Choo[2], Pauline M Rudd[1], Terry Nguyen-Khuong[1*]

1. Analytics Group, Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), Singapore 138668
2. Antibody Discovery Group, Bioprocessing Technology Institute, A*STAR, Singapore 138668
3. Waters Asia Pacific Pte Ltd, 1 Science Park Rd, #02-01/06 The Capricorn, Singapore Science Park II, Singapore 117528

***Corresponding Authors**

K. Wongtrakul-Kish: Email: k.wongtrakulkish@gmail.com

T. Nguyen-Khuong: Phone: (+65) 6407 4245. Fax (+65) 6478 9561. Email: terry_nguyen_khuong@bti.a-star.edu.sg

This document provides information supplemental to the main text, in the following sections:

1. Supporting Experimental Section
2. Supporting Figures

1. SUPPORTING EXPERIMENTAL SECTION

Extraction of GSLs

GSL extraction was performed based on Anugraham *et al*.[1]. Five mL of chloroform/methanol (2:1) was added to each cell pellet and left overnight at 4 ˚C on a spinning tube rotator. The samples were centrifuged at 1800 *g* for 20 min, and the supernatant was extracted. The pellet was re-extracted, and the supernatants were combined, following by drying under nitrogen gas. Crude GSLs were further purified by *n*-butanol/water partitioning according to Vidugiriene and Menon[2]. Dried GSLs were solubilized in 2 mL of n-butanol/water (1:1), vortexed, and centrifuged at 1000 *g* for 10 min. The upper butanol and lower aqueous layers were separated into individual glass vials. To the butanol layer, 1 mL of water/n-butanol (10:1) was added and mixed. To the lower aqueous layer, 1 mL of water/n-butanol (1:10) was added and mixed. Both mixtures were then subjected to centrifugation at 1000 *g* for 10 min. The combined butanol layers were dried under nitrogen gas.

HILIC-UPLC-FLR

Dried glycans and dextran were re-solubilised in 88 % acetonitrile/12 % water and separated at a temperature of 40 °C using an ACQUITY UPLC® BEH-Glycan column (1.7 µm, 2.1 x 150 mm). Gradient conditions were based on Albrecht *et al*.[3] as follows: 12 to 47 % (v/v) 50 mM ammonium formate pH 4.4 in acetonitrile at a flow rate of 0.56 mL/min from 0 - 36 min, followed by 47 to 70 % (v/v) at 0.25 mL/min from 39.5 to 42.0 min. LNFP1 and GM2 glycan were analysed at 30 °C with a flow rate of 0.4 mL/min and gradient conditions of 30 to 47 % (v/v) 50 mM ammonium formate pH 4.4 in acetonitrile from 0 - 34.8 min, followed by 47 to 80 % (v/v) from 34.8 to 36.0 min. The injection amounts were: 500 fmol for each GSL glycan standard, 7 % of breast cancer cell samples, and for GM2 glycan, the equivalent of 25 pmol of GM2 GSL was injected.

ESI-IM-MS

Samples were analysed in resolution mode and mobility separation performed in a traveling-wave drift tube. Spectra were acquired in positive ion mode with a full MS scan over a range of *m/z* 350-2000 and accumulation time of 1 s. The instrument conditions were as follows: 2.4 kV electrospray ionisation capillary voltage, 15 V cone voltage, 100 ˚C ion source temperature, 350 ˚C desolvation temperature, 850 L/hr desolvation gas flow, 40 L/hr cone gas flow, 650 m/s IMS T-wave velocity, and 40 V T-wave peak height. The T-wave mobility gas was nitrogen (N_2) and operated at a pressure of 3 mbar. The mobility cell was calibrated with Waters Major Mix IMS/Tof Calibration mix. Data acquisition was carried out using MassLynx™ (version 4.1).

Data Processing

1. Automated assignment library sets

Two library sets were used in the measurement of glycan assignment accuracy. First, to calculate overall glycan assignment accuracy, the attributes of 73 standard GSL glycans collected at Time-point 2 (compiled from six analyses) were matched to a multi-attribute library of the same standards constructed previously at Time-point 1 (compiled from eight analyses). To measure the assignment accuracy in distinguishing glycan monosaccharide linkages, the 73 GSL glycans were reduced to a subset library of 34 isomeric structures by removing structures with no isomers or structures that were compositional isomers (isobaric structures).

2. Automated glycan assignment

Correction factor: To account for drifts in CCS values between sample analysis days (Time-point 2) and CCS values generated for the construction of the library (Time-point 1), a correction factor was introduced to minimise the likelihood of such drifts impacting negatively on glycan matching accuracy. The CCS values of the dextran homopolymer that were analysed alongside unknown samples were compared to the CCS values of the dextran homopolymer run alongside the library standards and linear regression was used to determine the degree of change. Then, using the linear regression coefficients, the CCS values of sample glycans could then be aligned with those found in the library. As this step uses data from the independent dextran homopolymer standards and not the sample GSL glycans themselves, this does not unfairly bias the data processing.

Matching criteria: For automated glycan assignment, GU values, *m/z* and CCS values were extracted for each glycan peak and *m/z* used to pinpoint isomers. The GU values, *m/z* and CCS values were then searched against the multi-attribute library using Euclidean distance as the similarity measure. More precisely, given a glycan of unknown identity with n attributes, $U = \{u_1,..,u_n\}$, the distance between i^{th} library glycan, $G(i) = \{g_1^i, .., g_k^i\}$, and the unknown glycan can be computed only if $n = k$ as: $d^n(G(i),U) = \sqrt{\sum_{a=1}^{n}(u_a - g_a^i)^2}$ where u_a and g_a^i are the same type of attribute.

3. Benchmarking assignment accuracy and handling missing attributes

For an unknown test glycan, U, the minimum distance between U's attributes (generated after Time-point 2) and isomer attributes (generated at Time-point 1), was calculated as $d_{min}(\mathrm{U}) = \min\{d^n(G(1),U),..,d^n(G(73),U)\}$ where N is the number of identified isomers from *m/z* and $d_{min}(\mathrm{U})$ is a real number and is the criteria used to match to the library glycans. However, in some cases attributes were missing in

both the test GSL glycan and the library. To guarantee accuracy could be calculated for all 73 GSL glycans, the minimum distance $d_{min}(\mathrm{U})$, was calculated for all combinations of the five attributes (GU values, *m/z* and the three CCS values). For five test attributes observed (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+H]^{+}$,_ $^{TW}CCS_{N2}$ $[M+2H]^{2+}$, $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$), this involved calculating $d_{min}(\mathrm{U})$ in eight multi-dimensional libraries, namely: all possible combinations of library attributes (*m/z*, GU), (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+H]^{+}$), (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+2H]^{2+}$), (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$), (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+H]^{+}$, $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$), (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+2H]^{2+}$, $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$), (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+H]^{+}$, $^{TW}CCS_{N2}$ $[M+2H]^{2+}$), and (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+H]^{+}$, $^{TW}CCS_{N2}$ $[M+2H]^{2+}$, $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$). In the libraries where missing attributes occurred minimum distance could not be computed for a particular GSL glycan. Therefore, due to these incomputable distances the final annotation was the glycan corresponding to $d_{min}(\mathrm{U})$ that appeared in the majority of all eight libraries. When there were four test attributes observed, namely (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+H]^{+}$, $^{TW}CCS_{N2}$ $[M+2H]^{2+}$), (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+2H]^{2+}$,_ $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$), and (*m/z*, GU, $^{TW}CCS_{N2}$ $[M+H]^{+}$, $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$) there were four multi-dimensional libraries to consider (i.e. four possible combinations of attributes). For three test attributes, there were two multi-dimensional libraries to consider.

4. Accuracy as a function of Euclidean distance

Using the results from the assignment accuracies of the 73 GSL glycans (Timepoint 2 vs Time point 1), accuracy was computed as a function of distance $d^{n}(G(i),U)$ to calculate what level of $d^{n}(G(i),U)$ was required for high confidence in glycan assignment. Non-linear regression on the accuracy vs. distance $d^{n}(G(i),U)$ was used to estimate a probability of correct assignment for each possible attribute combination used in library matching.

5. Last resort computation

For cases where glycans were not found in the library, composition was given instead by permuting all possible GSL glycan compositions from the detected *m/z* values.

6. Statistics, clustering and visualization

To visualise the glycan attributes of GU, Mass, $^{TW}CCS_{N2}$ $[M+H]^{+}$, $^{TW}CCS_{N2}$ $[M+2H]^{2+}$ and $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$, a Principle Component Analysis was carried out. Pearson correlation coefficients were calculated using the R function 'corr.test'.

For breast cancer cell line profiling, all glycan assignments were confirmed manually and their probability of correct assignment calculated. Only glycans that were detected in two out of three replicates were kept for further analysis. For hierarchical clustering of breast cancer glycans, peak

areas were normalized using z-score[4] which standardizes the peak relative abundances to mean 0 and standard deviation 1. A hierarchy of clusters was built using the complete-linkage algorithm. Euclidean distance was used to calculate the dissimilarity among peaks. All p-values reported were found using a Student's paired t-test (assumes normal distribution).

2. SUPPORTING FIGURES

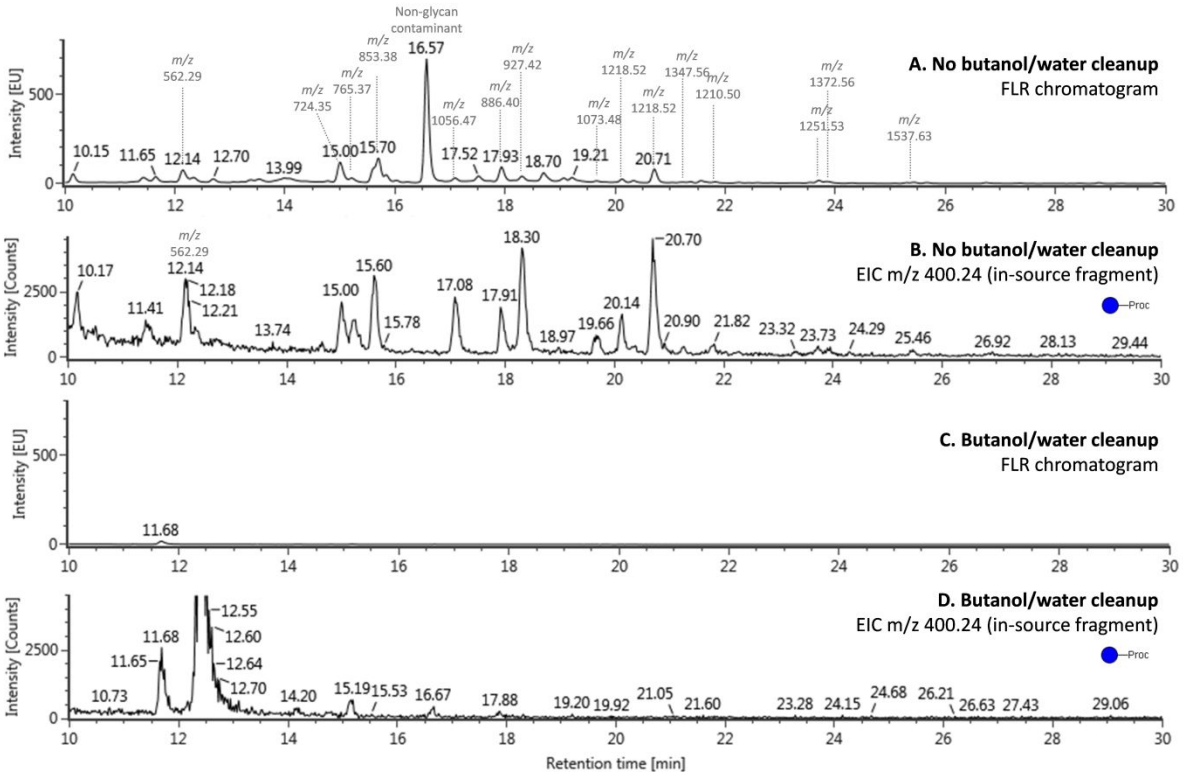
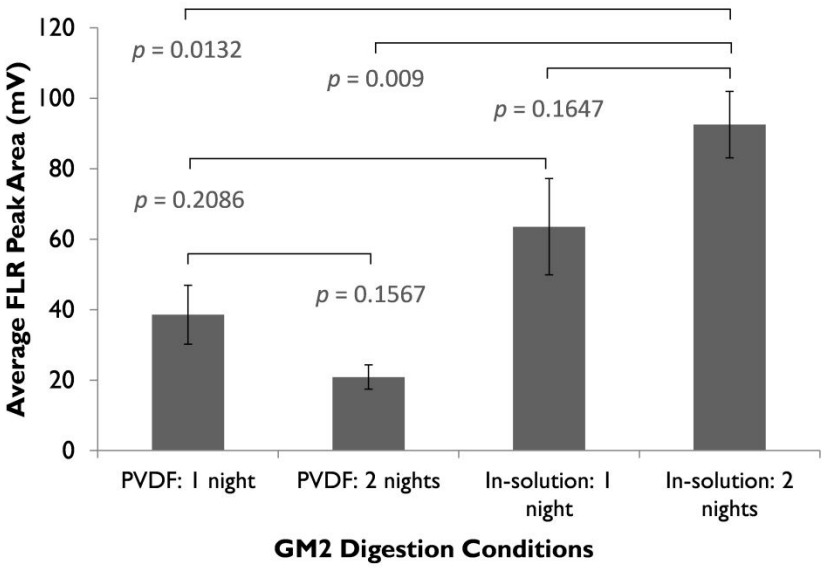


Figure S-1. Assessing butanol/water partitioning of GSLs extracted from BT474 breast cancer cells. The HILIC-FLD chromatogram of GSLs glycans (**A**) without prior butanol/water partitioning contained several potential glycan peaks. Peaks containing *m/z* values that correspond to glycan compositions are annotated. The (**B**) EIC of *m/z* 400.24 (an in-source fragment of the reducing end Glucose-Proc) was used to discriminate peaks containing true glycans and those containing non-glycan contaminants. To remove these non-glycan contaminant peaks, butanol/water partitioning of extracted GSLs from BT474 cells was performed prior to glycan release. The (**C**) HILIC-FLD chromatogram and (**D**) EIC of *m/z* 400.24 for these samples show that the partitioning step results in the loss of the majority of peaks.



147
148

149 **Figure S-2**. Assessment of GM2 glycan yield using HILIC-UPLC-FLR average peak areas. The highest
150 glycan yield was seen after two night's rEGCase II digestion performed in-solution and was significantly
151 higher than digestions performed on PVDF-bound GM2. Error bars denote standard deviation of the
152 mean.

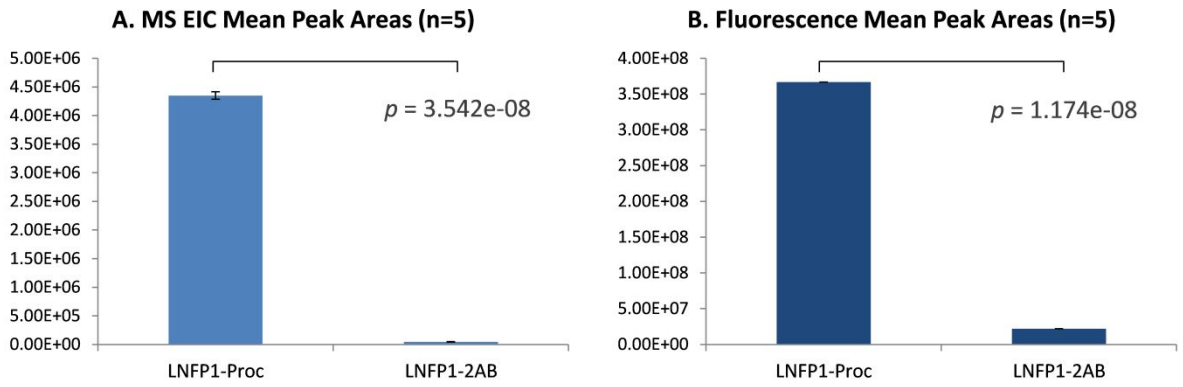


Figure S-3. (**A**) MS EIC and (**B**) FLD average peak areas were compared for procainamide and 2-AB labelled LNFP1 pentasaccharide. Average peak areas (n=5) were up to 16 times higher with FLR detection and 93 times higher with MS detection when using procainamide compared to 2AB. Error bars denote standard error of mean.

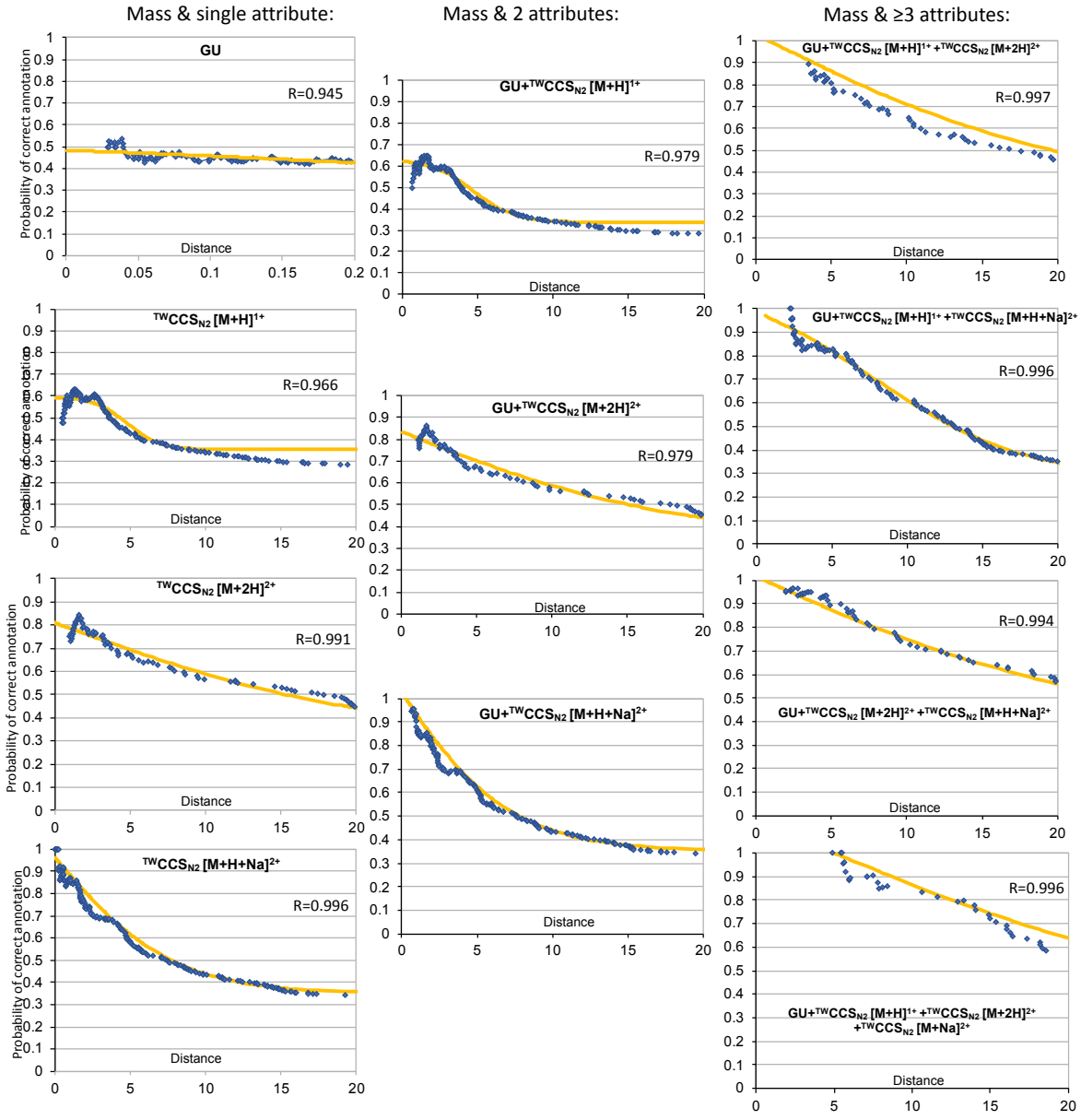


Figure S-4 Non-linear regression analysis of accuracy vs distance was conducted on different attribute combinations for the library GSL glycans. The blue points are estimated ratios of correct assignment given distance. The regression curves (orange) were used to calculate the probability of correct assignment for the GSL glycans identified in the breast cancer cell lines (Figure 4 and Table S-2). Depending on the attributes used to identify a particular glycan, the corresponding regression curve was used. R corresponds to coefficients of determination and show high correlation between distance and accuracy.

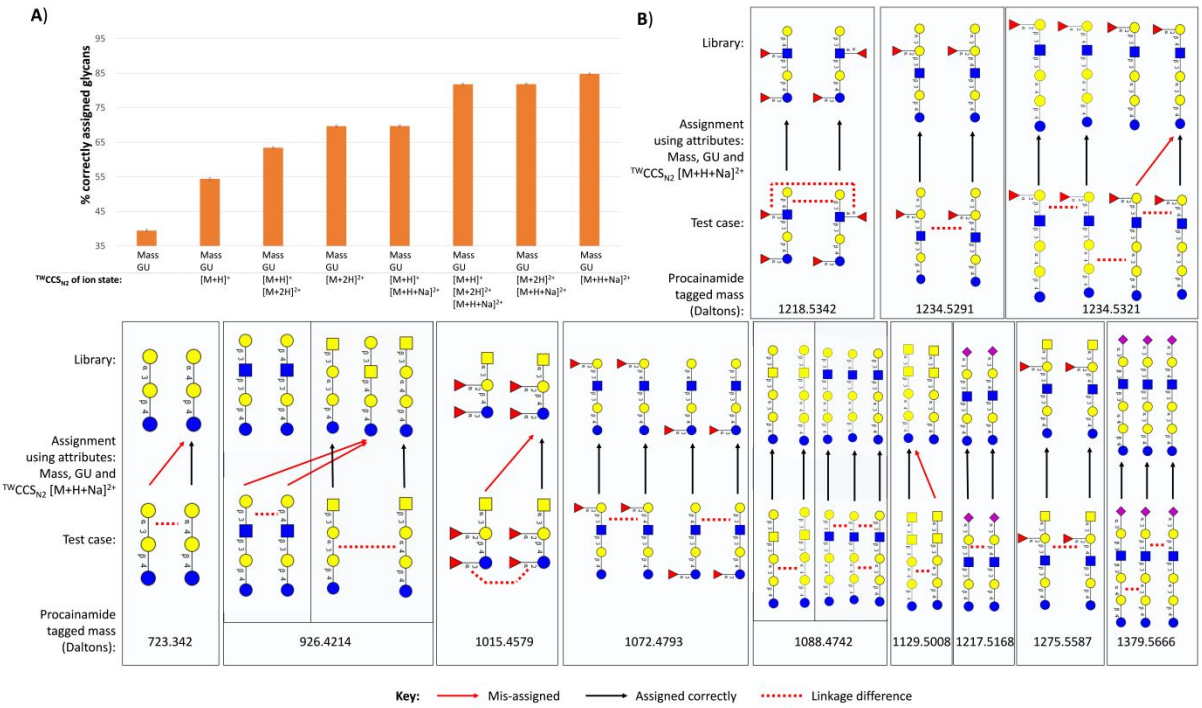


Figure S-5 Calculating the accuracy of mass and GU compared to multi-attribute-based glycan assignment in the differentiation of isomeric glycans (34 structures). **A**) Average assignment accuracies for all attribute combinations identified mass, GU and $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$ attributes to provide the highest accuracy (84.84 %). Averages and error bars were calculated by bootstrapping the 34 glycans. **B**) Visualisation of the 34 test cases used in this comparison (isomers grouped according to mass) and their library results when matched using mass, GU and $^{TW}CCS_{N2}$ $[M+H+Na]^{2+}$ showed incorrect assignments were not skewed towards particular linkage differences. Red arrows show mis-assigned glycans (6 out of 34) and black arrows show correctly assigned glycans. Dashed red lines show areas of monosaccharide linkage differences for an isomer group. Procainamide tagged masses correspond to those listed in Table S1.

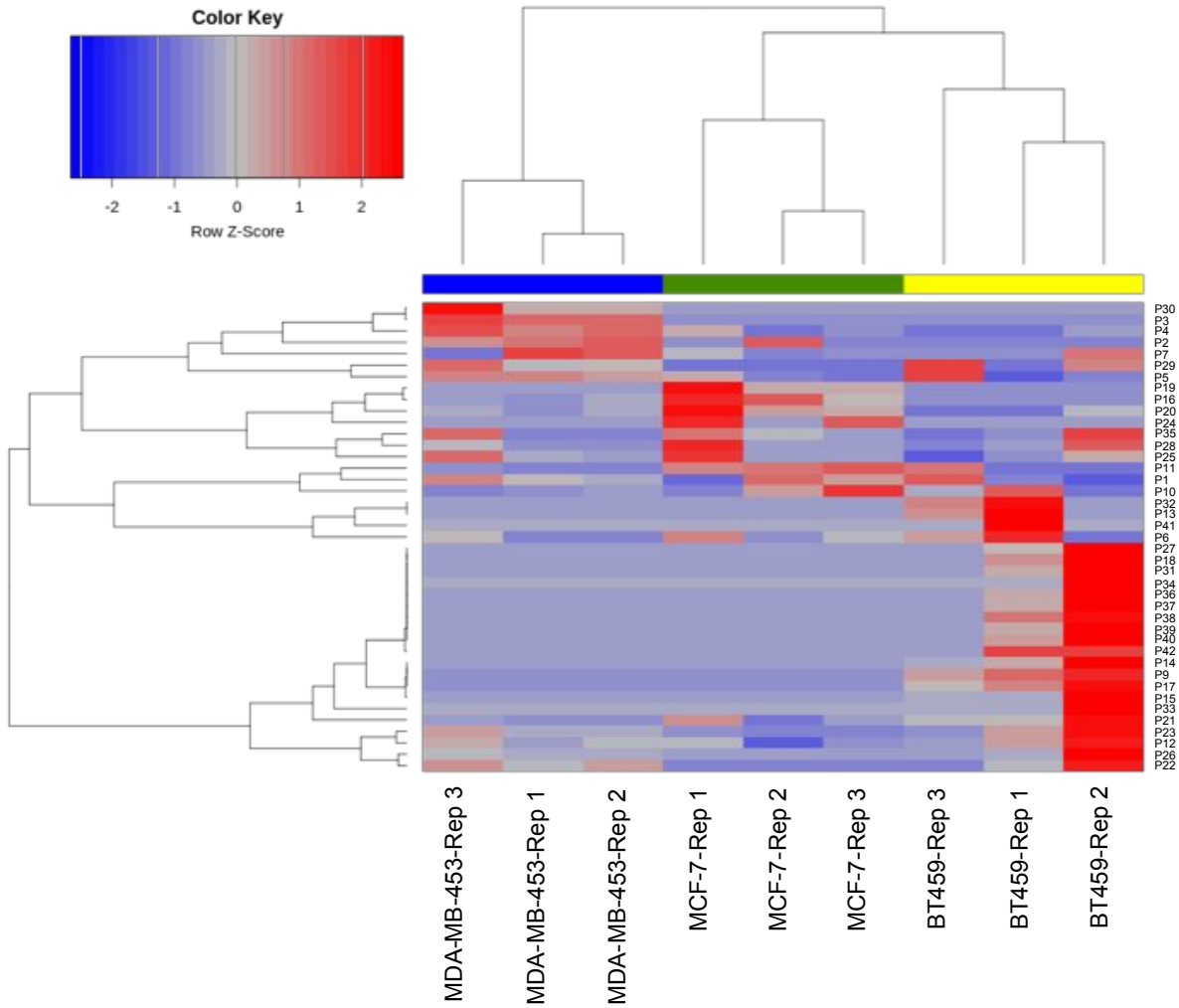


Figure S-6 Clustering analysis of LC-FLR peak average relative abundances of 33 peaks commonly detected in MDA-MB-453, MCF7 and BT459 cells analysed in triplicate. The analysis showed distinct glycosylation signatures for each cell line. Analysis was carried out in triplicate. Peak numbers correspond to those listed in Table S-3 and z-score denotes normalisation of the relative abundances to a mean equalling zero and standard deviation equalling one.

References

1. Anugraham, M.; Everest-Dass, A. V.; Jacob, F.; Packer, N. H., A platform for the structural characterization of glycans enzymatically released from glycosphingolipids extracted from tissue and cells. *Rapid Commun Mass Spectrom* **2015,** *29* (7), 545-61.
2. Vidugiriene, J.; Menon, A. K., Biosynthesis of glycosylphosphatidylinositol anchors. In *Methods in Enzymology*, Academic Press: 1995; Vol. 250, pp 513-535.
3. Albrecht, S.; Vainauskas, S.; Stockmann, H.; McManus, C.; Taron, C. H.; Rudd, P. M., Comprehensive Profiling of Glycosphingolipid Glycans Using a Novel Broad Specificity Endoglycoceramidase in a High-Throughput Workflow. *Anal Chem* **2016,** *88* (9), 4795-802.
4. Cheadle, C; Vawter, M. P.; Freed, W. J.; Becker, K. G., Analysis of Microarray Data Using Z Score Transformation. J Mol Diagnostics **2003**, *5* (2), pp 73-81