Supporting information

Analysis of N-glycoproteins using Genomic N-glycosite prediction

Shisheng Sun,[†] Bai Zhang,[†] Paul Aiyetan,[†] Jianying Zhou,[†] Punit Shah,[†] Weiming Yang,[†]

Douglas A. Levine,[‡] Zhen Zhang,[†] Daniel W. Chan[†] and Hui Zhang,[†]*

[†]Department of Pathology, Johns Hopkins University, Baltimore, Maryland 21231, USA.

[‡]Memorial Sloan-Kettering Cancer Center, New York, NY

Figure S1-4: Pages 6-9

Table S1: Matched N-glycopeptides from SKOV-3 cell line by GenoGlyco method with 10 ppm mass accuracy.

Table S2: N-glycopeptides identified from SKOV-3 cell line based on MS/MS fragmentation data and Sequest database search engine.

Table S3: N-glycopeptides identified in SKOV-3 cell line only by MS/MS spectra and database search engine.

Table S4: Matched N-glycopeptides from OVCAR-3 cell line by GenoGlyco method with 10 ppm mass accuracy.

Table S5: N-glycopeptides identified from OVCAR-3 cell line based on MS/MS fragmentation data and Sequest database search engine.

Table S6: N-glycopeptides identified from OCT-embedded tissues by GenoGlyco method.

Table S7: N-glycopeptides identified from OCT-embedded ovarian cancer tissue based on MS/MS database search.

EXPERIMENTAL SECTION

Deriving the cell expression Database. Microarray expression files to the NCI-60 cell-lines OVCAR-3, SKOV-3 were retrieved from the cell-miner repository developed by the NCI genomics and bioinformatics group[1]. Using the detection p-value estimation[2] implemented in the simpleaffy package on raw expression values, reliable probes were determined. Reliable probes were annotated using the hgu133plus3.db package to extract respective Uniprot accession of statistically significant detected transcripts. Detected corresponding protein sequences in the each cell line were extracted from the IPI human protein database (v3.87)[3] and placed in separate fasta databases. The scripting and analysis were done in the R/Bioconductor programming environment[4]

The cell specific *N*-glycopeptide candidate databases. Potential *N*-glycopeptides of cells were predicted from the cell expression database by the following steps: Firstly, the "N" in all N-X-S/T motifs (where X cannot be proline) was replaced by "n" to mark as potential *N*-glycosylation sites. Then, In silico tryptic digestion mimics of proteins with up to two missed cleavages were performed as the trypsin cleavage patterns in biological experiments, which cleaves peptide chains at the carboxyl side of the amino acids lysine (K) or arginine (R), except when either is followed by proline (P). All Peptides containing "n" were regarded as potential *N*-glycopeptides. Finally, the monoisotopic mass of each potential *N*-glycopeptide was calculated with carbamidomethytion (C), oxidation (M) and deamidation (N) at potential N-glycosylation sites (n) as fixed modifications. The cell line specific *N*-glycopeptide candidate databases for both SKOV-3 and OVCAR-3 cell line were then created based on these

prediction results. We developed a customized, open source Python program to implement the above analysis pipeline (available upon request).

Feasibility Analysis of mass mapping of SILAC peak pair method. The theoretical feasibility of the SILAC peak pair method for identification of *N*-glycopeptides was investigated by distinguishing each potential *N*-glycopeptide with others from same cell line using both mass and K/R count information in peak pairs with 1-10ppm mass error tolerances. The number of peptides at each peptide mass within certain mass error tolerance was counted and the percentage of each matched peptide count was used to demonstrate the feasibility of the method. If the mass and K/R count information of each peptide (contained in SILAC peak pair) are sufficient to distinguish the majority of *N*-glycopeptides in the database, it will demonstrate, at least theoretically, that *N*-glycopeptide identification of the cell line can be achieved by SILAC peak pair at MS1 level. The scripts here and below were written and run in R environment [1].

REFERENCES

1. Shankavaram, U. T.; Reinhold, W. C.; Nishizuka, S.; Major, S.; Morita, D.; Chary, K. K.; Reimers, M. A.; Scherf, U.; Kahn, A.; Dolginow, D.; Cossman, J.; Kaldjian, E. P.; Scudiero, D. A.; Petricoin, E.; Liotta, L.; Lee, J. K.; Weinstein, J. N., *Mol. Cancer Ther.* **2007,** *6*, 820-832.

2. Liu, W.-m.; Mei, R.; Di, X.; Ryder, T. B.; Hubbell, E.; Dee, S.; Webster, T. A.; Harrington, C. A.; Ho, M.-h.; Baid, J.; Smeekens, S. P., *Bioinformatics* **2002,** *18*, 1593-1599.

3. Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R.,

Proteomics **2004,** *4*, 1985-1988.

4. Gentleman, R.; Carey, V.; Bates, D.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J.; Zhang, J., *Genome Biol.* **2004,** *5*, R80.

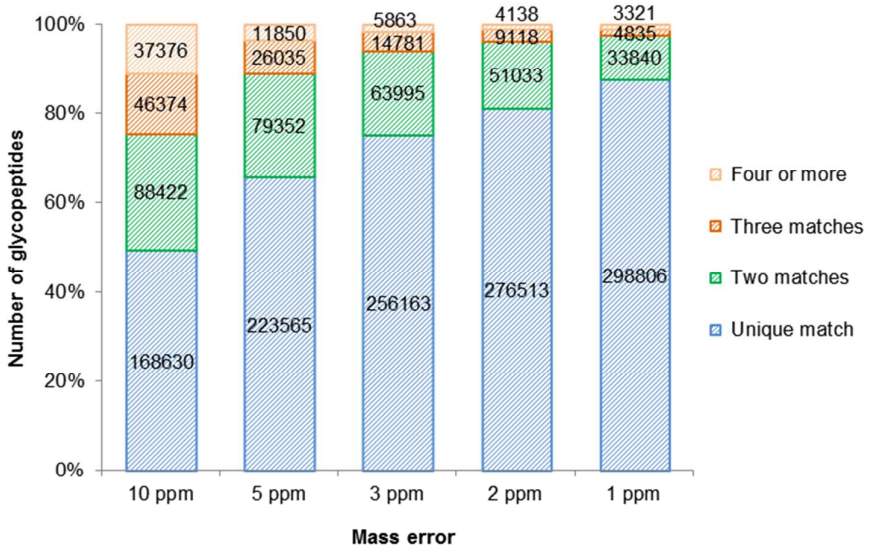

Figure S1. The theoretical feasibility of the glycoprotein identification in SKOV-3 cell line using GenoGlyco method and RefSeq entire database. The numbers represent the number of *N*-glycopeptides with distinct mass with K/R count information in OVCAR-3 cell line. Deamidation was set as variable modification from the second potential *N*-glycosite of a peptide if the peptide contains more than one N-X-S/T motif.

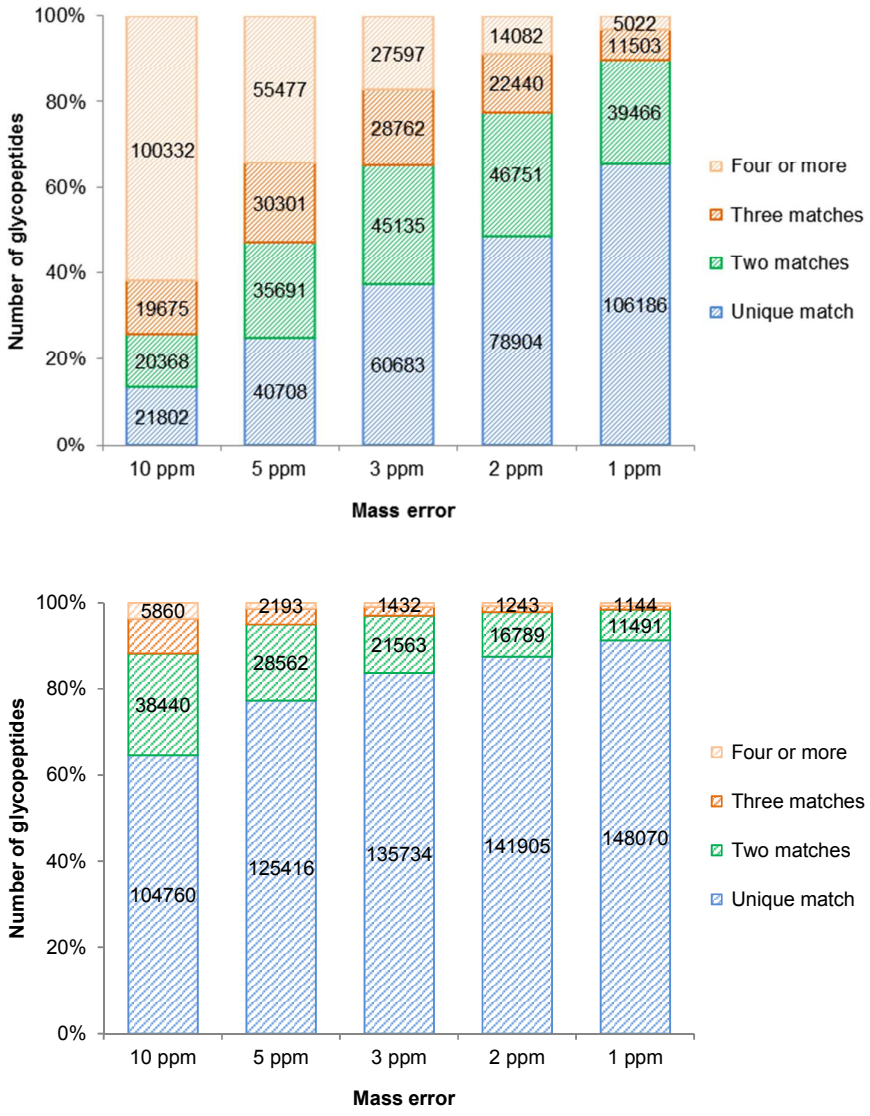

Figure S2. The theoretical feasibility of the glycoprotein identification in OVCAR-3 cell line using GenoGlyco method. The numbers represent the number of *N*-glycopeptides with distinct mass with (below) or without K/R count information (upper) in OVCAR-3 cell line. The N-glycopeptides were predicted from OVCAR-3 cell expressed database.

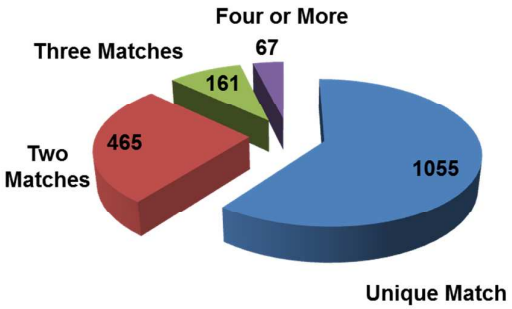

Figure S3. The performance analysis of N-glycopeptides isolated from OVCAR-3 cells using GenoGlyco method.

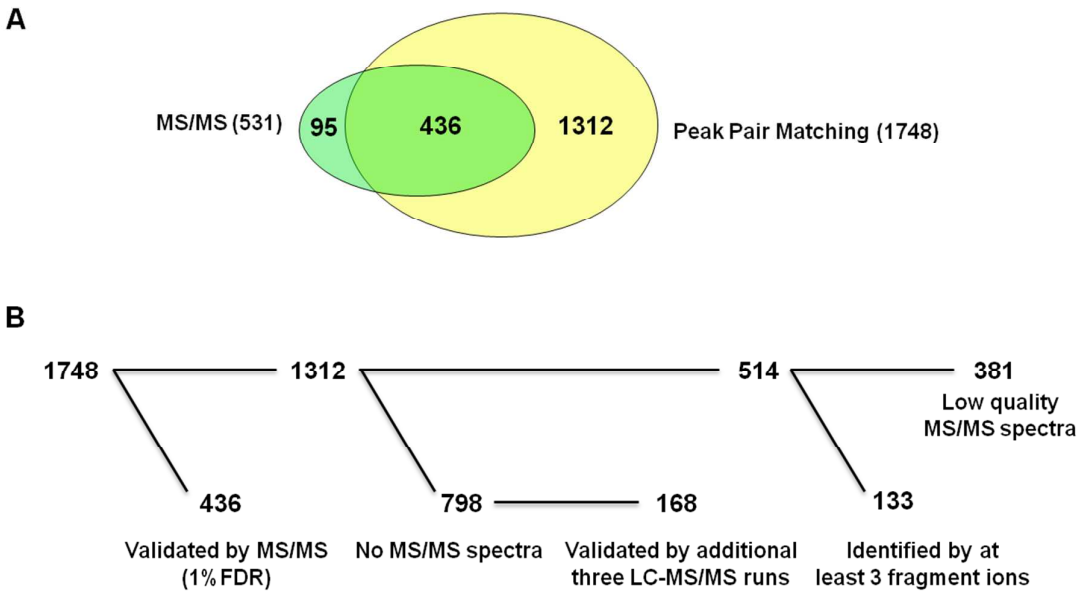

Figure S4. Analysis of N-glycopeptides identified in OVCAR-3 cell line by GenoGlyco and MS/MS based identification. (A) The comparison of N-glycopeptides identified between using GenoGlyco method and LC-MS/MS and database search. (B) Examination of identified 1,748 N-glycosites by GenoGlyco. 436 N-glycosites were verified by MS/MS identification, the remaining 1,312 of N-glycosites, 798 were not subjected for data dependent MS/MS analysis during the same LC-MS-MS/MS analysis, but additional 3 LC-MS/MS runs of the light N-glycopeptide samples identified and verified 168 of these N-glycosites. There were MS/MS spectra generated for the rest 514 SILAC labeled glycopeptide pairs, 133 glycosites contained MS/MS spectra that contained at least 3 fragment ions matched to theoretical fragment ions using targeted spectrum investigation, and the remaining 381 assigned glycopeptides contained low quality of MS/MS spectra.