THE EMBO JOURNAL

Manuscript EMBO-2012-84257

Precision Mapping of the Human O-GalNAc Glycoproteome through Human SimpleCell Technology

Catharina Steentoft, Sergey Y Vakhrushev, Hiren J Joshi, Yun Kong, Malene B Vester-Christensen, Katrine T-B. G Schjoldager, Kirstine Lavrsen, Sally Dabelsteen, Nis B Pedersen, Lara Marcos-Silva, Ramneek Gupta, Eric paul Bennett, Ulla Mandel, S¯ren Brunak, Hans H Wandall, Steven B Levery, Henrik Clausen

Corresponding author: Henrik Clausen, Copenhagen Center for Glycomics

Review timeline:

Submission date:	18 December 2012
Editorial Decision:	04 February 2013
Revision received:	15 February 2013
Editorial Decision:	26 February 2013
Additional correspondence (author):	04 March 2013
Revision received:	15 March 2013
Accepted:	18 March 2013

Editor: Hartmut Vodermaier

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 04 February 2013

Thank you again for submitting your manuscript on comprehensive O-GalNAc glycoproteomics for consideration by The EMBO Journal, and please excuse the delay in its evaluation associated with the end-of-the-year holiday break. We have now received the comments from three expert referees (copied below), and I am pleased to inform you that all of these referees find this work of interest and potential importance. We should therefore be happy to consider the study further for publication, pending adequate revision in response to the various points raised by the reviewers. As you will see, the majority of the concerns pertain to issues of interpretation, discussion and explanation, and related to that to data analysis and presentation, and should therefore be straightforward to address. Nevertheless there are also a limited number of points that will require some additional experimentation, most notably referee 1's point 1 and referee 2's point 2.

I would therefore like to invite you to revise the manuscript according to the referee reports, and to resubmit it together with a thorough and diligent point-by-point response letter. Please bear in mind that this letter will form part of the Peer Review Process File available online to our readers in the case of publication (for more details on our Transparent Editorial Process initiative, please visit our website: http://www.nature.com/emboj/about/process.html).

When revising the manuscript, I would encourage you to also modify the title according to referee 3's suggestion, to change "GalNAc-Type O-Glycoproteome" into "O-GalNAc Glycoproteome". I am also wondering if the somewhat unwieldy (and in the case of 'Human' also repetitive) phrase "Enabled by Human SimpleCells" could be simplified into something like "through SimpleCell technology"? Any alternative proposals would of course also be welcome.

We generally allow three months as standard revision time, and it is our policy that competing manuscripts published during this period will have no negative impact on our final assessment of your revised study. However, we request that you contact the editor as soon as possible upon publication of any related work, to discuss how to proceed. Should you foresee a problem in meeting this three-month deadline, please let us know in advance and we may be able to grant an extension.

Thank you for the opportunity to consider this work for publication. I look forward to your revision.

REFEREE REPORTS:

Referee #1

In the manuscript entitled "Precision Mapping of the Human GalNAc-Type O-Glycoproteome Enabled by Human SimpleCells", the authors are applying the SimpleCells strategy to 12 human cancer cell lines provided a global view of the human O-glycoproteome with more than 600 O-glycoproteins and almost 3,000 glycosites, and also present an improved NetOGlyc4.0 model for prediction of O- glycosylation. This reviewer believes the experiments and the data presented mostly support the conclusions, and the manuscript should be considered for publication on EMBO J after the authors addressed the following concerns.

Concerns

1) In this manuscript, the authors have presented a large dataset of O-Glycosites without natural O-glycan structure information. The authors need to provide some of O-glycan structure information by other strategy, at least some of the identified proteins at the validation step. It is very important to exploration of how site-specific O-glycosylation regulates protein function.

2) On page 6 and page 11, the authors found "almost 50% of the identified glycoproteins and glycosites were found only in one cell line, and each cell line contributed a number of unique glycoproteins", and the authors concluded: "The finding of unique subsets of O-glycoproteins in each cell line provides evidence that the O-glycoproteome is differentially regulated and dynamic" in the abstract. It might be not appropriate. Firstly, the enrichment might not be that perfectly efficient for all O-glycoproteins and O-glycopeptides by lectin weak affinity chromatography with Vicia villosa agglutinin (LWAC-VVA), even though it was two-stage enrichment at both protein and peptide level. Secondly, the authors used trypsin and chymotrypsin for protein digestion, and the over-all sequence coverage might not be sufficient in the MS analysis (both HCD-MS2 and ETD-MS2). Even though the authors claimed "in a similar analysis of the global proteome of 11 human cell lines it was found that almost 90% of the identified proteins were found in all cell lines, albeit in different levels (Geiger et al., 2012)". The review believed that the sequence coverage and enrichment efficiency might not be comprehensively to cover all real O-glycosites. Thirdly, the O-glycosylation is highly dynamic, and the cycle time might be in a very wide range even in the same cell line. Hence, the conclusion might be not entirely appropriate.

3) In the Extended Experimental Procedures section, the authors need to provide more information regarding the Mass Spectrometry and Data Analysis parameters. For example, the parameters of Dynamic Exclusion for LC-MS/MS; the parameters of ETD Reaction Time for ETD-MS2; the signal intensity threshold for extraction mass of 1x, 2x, 3x, and 4x HexNAc to distinct mgf files.

Referee #2

Steentoft et al have used their method of zinc-finger deletion of the Cosmc chaperone for generating 'simple cells' that do not extend the O-glycans and thus allowing isolation of such peptides and characterization of the O-glycosite. They have now used this approach to 10 cell lines and by this extended the O-glycositeome considerably. The results show that this is much larger than appreciated previously and for example the localization of a site in the LDLR A-folds might have large also medical implications.

This manuscript is an important step forward in our understanding of the O-gycositeome and opens a new research arena. Although previously realized by people in the field, the authors point out that

certain regions (stakes, in between domains) are more likely to the O-glycosylated. The manuscript is well written and in addition to pointing out the future, also specific information for certain proteins.

MAJOR
1. The most important message is that it is the combination of the protein to be O-glycosylated and one or several out of the 20 GalNAcTs that are necessary for a site-specific O-glycosylation. The NetOGlyc database was developed at a time when it was assumed to exists only ONE GalNAcT and thus the approach of one unifying algorithm was correct. Unfortunately, the NetOGlyc has misled many scientist that have not been sufficiently well acquainted with glycobiology. Thus it is surprising that the authors has chosen to just release another version of NetOGlyc, although they have incorporated novel features like domain predictions that has improved its predictive capability. That it is impossible to make a useful tool along the idea of one GalNAcT should be pointed out strongly to be understood for the nonexpert. The focus on NetOGlyc lowered. The authors are encouraged to instead start over with a tool designed to address the true reality as outlined in this manuscript and probably presented in their GlycoDomainViewer (unfortunately, reviewers were not given information to access this site).

2. Although, as pointed out by the authors (p.7), it is premature to make a positive correlation of GalNAcTs and glycosite, the opposite is possible. The authors should provide a complete list of expressed GalNAcTs in these 10 cell lines. Half the GalNAcTs have been analyzed by immunhistochemistry, but probably due to lack of antibodies the other 10 are missing. It should be easy for this competent group to also analyze the presence of all the GalNAcTs using RT-PCR. This will allow others to at least exclude some GalNAcTs from attaching GalNAc to the sites found in the specific cell line.

3. Unfortunatley, 'unstructured' regions rich in the amino acids Pro, Thr and Ser have so far not got a common name in databases like PFAM etc. Lang et al (PNAS (2007) 104, 16209) introduced the term PTS (-domain, -sequence, -region) for such sequences that become O-glycosylated and when long generate a mucin domain. The authors are encouraged to use nomenclature especially as they use the same letters, but in another order. Alter p. 7, STP, and p. 12, PST two times into PTS. Also rephrase and add PTS to the 'tandem repeat region' page 7, l. 9. Some PTS regions have VNTR, but not all.

MINOR
1. p. 5, l. 18: Alter breadth to depth.

2. p. 7-8: Should have been very helpful for the community to pull out the 14% proteins that were classified as nuclear into a separate supplementary Table. Among these are probably significant proteins taking part in signaling.

Referee #3

This manuscript presents a first look at the human O-GalNAc proteome and is of broad interest to biological scientists. Using a ZFN strategy to remove COSMC and therefore C1GALT1 activity in 12 human cell lines, the authors purified glycopeptides with terminal GalNAc residues using weak lectin affinity chromatography and performed nLC/MS/MS to identify close to 3000 glycosites and develop a new glycosite predictor NetOGlyc 4.0. This represents an ~10-fold increase in currently existing glycosites and is a major leap forward. The manuscript is quite clearly written, though it would benefit from fuller explanations of certain points, and corrections as follows:

1) Why is the existence of Tyr-O-GalNAc referred to as "likely" in abstract? Were the calls ambiguous?

2) The qualifications relating to the general strategy of isolating O-GalNAc peptides should be spelled out in the first section of the Results and Discussion. For example, the addition of Gal to form the T antigen on a glycoprotein may promote or inhibit the transfer of GalNAc to a "nearby" Ser/Thr; dense regions of O-GalNAc are/may be resistant to proteolysis; O-GlcNAc on cytosolic regions or contaminating proteins or on glycoproteins in the secretory pathway with EGF repeats containing a particular consensus sequence (defined in (Alfaro et al (2012) PNAS 109, 7280))

cannot be distinguished from O-GalNAc by mass; only O-GalNAc glycopeptides that bind to VVA will be detected.

3) Were O-GalNAc glycopeptides that did not bind to the VVA column during LWAC analyzed by MS/MS? If so, the results should be summarized.

4) The O-GlcNAc question should be explicitly addressed in the main text. In Methods it is mentioned that glycopeptides in cytosolic domains are likely to represent O-GlcNAc. However, it is now clear that glycoproteins in the secretory pathway other than Notch receptors carry O-GlcNAc (Alfaro et al (2012) PNAS 109, 7280). This reference should be quoted and discussed.

5) Are glycopeptides from glycoproteins with no signal peptide actually cytosolic, O-GlcNAcylated proteins? They should be compared to glycopeptides previously identified in O-GlcNAc databases, especially the recent analysis by Alfaro et al (PNAS (2012) 109, 7280).

6) Validation of the superior prediction capabilities of NetOGlyc v4.0 should be spelled out by identifying O-GalNAc glycosites predicted by v4.0 but not v3.0, and actually identified by MS/MS in this or other studies.

7) The description of the GlycoDomain Viewer should be clarified. For example, the following sentences in Results are confusing:
"Using the GlycoDomainViewer we determined that over 80% of the sites occurred outside the curated set of domains. In contrast, N-glycosylation motifs appeared within domains in 78% of the referenced annotated sites in Uniprot of our dataset."
---80% of which sites? The N-glycosylation motif is not defined. Were the broadest N-glycosylation motifs used according to the most recent proposals?

8) An expanded legend to Fig. 5 explaining the diagram is necessary. How were N-glycan sites identified? How were N-glyosites predicted? How were O-glycosites predicted? Better to use N rather than a chitobiose core symbol for N-glycans. Identified N-sites could be bolded N.

9) Predicted O-GlcNAc modifications if any, should be included in Fig. 5.

10) The number of v3.1 glycosites in Fig. 6C is extremely high. Appears to be an error.

11) The legend to Fig. 7 should be expanded to clarify the figure. Fig. 7 itself is hard to read. Bolding the T enzymes may help.

12) The in vitro assay used on 181 peptides represents a very crude comparison as noted by the authors. The Discussion should briefly present how to go about this comparison to obtain better answers in a larger studyi.

13) The manuscript should specifically address the O-glycosite/GalNAcT isoform consensus sequences previously proposed by Gherken and others in relation to the ~3000 glycosite sequences reported here. When all available data are taken into account, does a common consensus sequence emerge? Do the 181 peptide substrates fit predictions?

Minor points:
a) The abbreviations used in Fig. 4 (ext. cel. -- etc.) are not intuitive or usual.
b) The Gill review is 2010 in bibliography -- should be 2011.
c) Sentence needs changing: GalNAc-T isoforms differs in cells and tissues and changes during - delete second "and"
d) The Discussion suggests that this paper introduces the SC strategy and should be modified.
e) Table S3 needs a legend to describe mAb row and the numbers after each cell line.
f) "GalNAc-Type O-Glycoproteome" is unwieldy in the title. Suggest "O-GalNAc Glycoproteome" as a more specific and enduring title.

1st Revision - authors' response 15 February 2013

Re. EMBOJ-2012-84257
Revised manuscript submission Feb. 14, 2013

Point-by-point list of responses to queries:

Referee #1
Query 1) In this manuscript, the authors have presented a large dataset of O-Glycosites without natural O-glycan structure information. The authors need to provide some of O-glycan structure information by other strategy, at least some of the identified proteins at the validation step. It is very important to exploration of how site-specific O-glycosylation regulates protein function.

Answer 1.
We agree with the reviewer that O-glycan structure information indeed is important for evaluating biological functions of O-glycosylation. Considerable information of O-glycan structures derived from glycoprofiling of cells and tissues already exist in the literature. The focus of our study is the sites of O-glycans and the technology we developed using SimpleCells eliminates O-glycan extension. Thus, we demonstrate that all SimpleCell lines lack core 1 structures and express Tn (GalNAc), STn (NeuAc-GalNAc), or a combination in Suppl. Fig 2. Currently there are no methods available to analyze O-glycan structures and sites together on a proteome-wide scale, and we envision that the data presented in the study eventually will enable development of proteome-wide analysis using precursor ion selection for known sites as concluded in the Discussion.
We have previously shown that the O-glycoproteome (in terms of sites) found with the SimpleCell technology correlates very well with the O-glycoproteome found in wildtype cells (Steentoft et al. Nature Methods 2011 and referenced). This was possible with the K562 cells because they produce a rather homogenous core1 O-glycosylation, but this is not applicable to other cells. We have also previously demonstrated the validity of our O-glycoproteome data with two model proteins, ApoC-III and ANGPTL3 (Schjoldager et al PNAS 2012 and referenced).

Action: We have expanded the Introduction (last paragraph) to clarify previous validation of the SimpleCell strategy as follows:
"We previously applied this strategy to a few human cell lines and demonstrated efficient identification of O-glycoproteins and sites of O-glycosylation. We showed that the O-glycoproteome identified in K562 SimpleCells was essentially identical to that found in wildtype K562 cells using a similar strategy of analysis (Steentoft et al., 2011). We also applied the SimpleCell strategy to pinpoint non-redundant O-glycosylation performed by a single polypeptide GalNAc-T using differential analysis of O-glycoproteomes produced in an isogenic cell model with and without knockout of one GalNAc-T isoform (Schjoldager et al., 2012)."

Query 2) On page 6 and page 11, the authors found "almost 50% of the identified glycoproteins and glycosites were found only in one cell line, and each cell line contributed a number of unique glycoproteins", and the authors concluded: "The finding of unique subsets of O-glycoproteins in each cell line provides evidence that the O-glycoproteome is differentially regulated and dynamic" in the abstract. It might be not appropriate. Firstly, the enrichment might not be that perfectly efficient for all O-glycoproteins and O-glycopeptides by lectin weak affinity chromatography with Vicia villosa agglutinin (LWAC-VVA), even though it was two-stage enrichment at both protein and peptide level. Secondly, the authors used trypsin and chymotrypsin for protein digestion, and the over-all sequence coverage might not be sufficient in the MS analysis (both HCD-MS2 and ETD-MS2). Even though the authors claimed "in a similar analysis of the global proteome of 11 human cell lines it was found that almost 90% of the identified proteins were found in all cell lines, albeit in different levels (Geiger et al., 2012)". The review believed that the sequence coverage and enrichment efficiency might not be comprehensively to cover all real O-glycosites. Thirdly, the O-glycosylation is highly dynamic, and the cycle time might be in a very wide range even in the same cell line. Hence, the conclusion might be not entirely appropriate.

Answer 2.
We fully agree with the reviewer that the analysis performed is not all-inclusive and this has been stated several places in the text as well as the title. This is a first assessment of the O-glycoproteome and sites of O-glycan attachment using one analytical strategy, and we do provide ideas for future directions in the Discussion aimed at expanding the depth of analysis. However, as discussed under address to comment #1, we have demonstrated the validity of comparing glycoproteomes obtained from different cells using the same digestion and isolation strategy in Schjoldager et al PNAS 2012. While such comparisons given the nature of mass spectrometry are never 100%, the large differences in O-glycoproteomes found among cell lines clearly suggest unique subsets of O-glycoproteins in cells. This is of course further substantiated by the differential expression of GalNAc-Ts in cells and our study showing that e.g. ApoC-III and ANGPTL3 are O-glycoproteins in HepG2 both not in HepG2 without GalNAc-T2 (Schjoldager et al PNAS 2012).
The reviewer appear to state that the O-glycoproteome is highly dynamic (in thirdly) as if this is generally accepted, however, there is to our knowledge really no data to support this and it remains a good hypothesis. Regardless, we fail to see how this could affect the outcome sum of the applied O-glycoproteomics strategy with standardized 72hrs growth of cloned cell cultures?

Action: We have further stressed these points by including the following sentence in the end of Discussion:
"This is a first view of the human O-glycoproteome and we anticipate further expansion through use of additional cell lines, different protease digestion strategies, and more sensitive instrumentation."

Query 3) In the Extended Experimental Procedures section, the authors need to provide more information regarding the Mass Spectrometry and Data Analysis parameters. For example, the parameters of Dynamic Exclusion for LC-MS/MS; the parameters of ETD Reaction Time for ETD-MS2; the signal intensity threshold for extraction mass of 1x, 2x, 3x, and 4x HexNAc to distinct mgf files.

Answer 3.

We agree.

Action: The following text has been added (Supplemental Page 7, end of Par. 1, Lines 14-20):

"In general, activation times were 30 msec and 100-200 msec for HCD and ETD fragmentation, respectively; isolation width was 4 mass units, and usually 1 microscan was collected for each spectrum. Automatic gain control (AGC) targets were 100,000 ions for Orbitrap MS1 and 10,000 for MS2 scans, and the AGC for fluoranthene ion used for ETD was 300,000. Supplemental activation (20%) of the charge-reduced species was used in the ETD analysis to improve fragmentation. Dynamic exclusion for 30 sec was used to prevent repeated analysis of the same components. Polysiloxane ions at *m/z* 445.12003 were used as a lock mass in all runs."

and also, to the following sentence (Supplemental Page 7, Par. 2, lines 7-10), the text noted here in **bold highlight** has been added:

"As a further pre-processing procedure, all HCD data showing the presence of fragment ions at *m/z* 204.08 were extracted into a single .mgf file **(signal intensity threshold, 1.5)**, and the exact mass of 1x, 2x, 3x, and 4x HexNAc units was subtracted from the corresponding precursor ion mass, generating four distinct files."

Referee #2
MAJOR
Query 1. The most important message is that it is the combination of the protein to be O-glycosylated and one or several out of the 20 GalNAcTs that are necessary for a site-specific O-glycosylation. The NetOGlyc database was developed at a time when it was assumed to exists only ONE GalNAcT and thus the approach of one unifying algorithm was correct. Unfortunately, the NetOGlyc has misled many scientist that have not been sufficiently well acquainted with glycobiology. Thus it is surprising that the authors has chosen to just release another version of NetOGlyc, although they have incorporated novel features like domain predictions that has improved its predictive capability. That it is impossible to make a useful tool along the idea of one GalNAcT should be pointed out strongly to be understood for the nonexpert. The focus on NetOGlyc lowered. The authors are encouraged to instead start over with a tool designed to address the true reality as outlined in this manuscript and probably presented in their GlycoDomainViewer (unfortunately, reviewers were not given information to access this site).

Answer 1.
Well, this is a blunt and partially correct statement that we have ourselves dealt with for years!! Note that NetOGlyc was in fact developed in 1998 and already in 1996/7 was it clear that multiple GalNAc-T isoforms existed (see eg Wandall et al JBC 1997). The fundamental question though is if it is indeed mutually exclusive to have/identify global and isoform specific motifs? While the former may be broader more diffuse motifs (algorithms) than isoform specific, we have to realize that the general user will have much more use of a global predictor since knowledge of the GalNAc-T repertoire in play for a particular protein is likely to be unknown. We are not yet in position to have enough isoform specific data to make meaningful isoform specific predictors, so this is currently not an option. However, why is a global algorithm necessarily misleading? The GalNAc-Ts controlling initiation and sites of O-glycosylation are highly homologous and although they have some distinct functions there is a great overlap especially among the workhorses, GalNAc-T1-4, as also evidenced by the current data. It is thus quite reasonable that general features of O-glycosylation sites exists that can serve as basis for a global predictor. Having said this the NetOGlyc3.1 has extremely low predictive power for especially all the isolated O-glycosylation sites identified for the first time in this study. The new NetOGlyc4.0 is not just another version but an entirely new algorithm for the first time founded on a (relatively) non-biased and much larger dataset. Interestingly, as shown in Fig. 6 and

pointed out in the text, the two algorithms identify the fairly similar subset of the proteome as O-glycoproteins while it is the predictions of actual sites that vary tremendously.
It is unfortunate that instructions for access to our GlycoDomainViewer did not reach the reviewers for some reason, but the current predictor is available at http://glycodomain.glycocode.com with username:embo and password:review.
We chose not to dwell at length on the above in the text and included the main point as follows in the Discussion:
"The current predictor is based on a unifying model for all GalNAc-Ts, but improved results may be obtained by collecting sufficient isoform-specific data to produce isoform-specific algorithms."

Action: We provide the reviewers access to our GlycoDomainViewer, which displays direct comparison of NetOGlyc3.1 and 4.0 predictions.

Query 2. Although, as pointed out by the authors (p.7), it is premature to make a positive correlation of GalNAcTs and glycosite, the opposite is possible. The authors should provide a complete list of expressed GalNAcTs in these 10 cell lines. Half the GalNAcTs have been analyzed by immunohistochemistry, but probably due to lack of antibodies the other 10 are missing. It should be easy for this competent group to also analyze the presence of all the GalNAcTs using RT-PCR. This will allow others to at least exclude some GalNAcTs from attaching GalNAc to the sites found in the specific cell line.

Answer 2.
While we agree that the end goal is to be able to correlate GalNAc-T repertoire with the O-glycoproteome, the premise for this point is incorrect. It is indeed premature to look for both positive and negative correlations because GalNAc-T1 and T2 are expressed in all cell lines tested. Nevertheless, it is important to characterize the repertoire of enzymes. Our lab has for almost 20 years refrained from use of semi-quantitative RT-PCR or qPCR as these really only can be used for comparative analysis of levels (of transcripts) and not for assessing enzyme protein levels or even for simple presence or absence evaluation (what are the cutoff levels?). We strongly believe that these data would only confuse things further, and given our (huge) efforts to establish a panel of Mabs to define expression levels and intracellular topology of the GalNAc-T family, we would hate to muddy the field with some semi-quantitative PCR data at this stage. Also it is no small undertaking to develop either qPCR or RNASEQ data for all 20 genes when starting from scratch?
We may also point out that we have previously made a comprehensive literature comparison of organ expression of GalNAc-Ts evaluated by different strategies including Mabs, PCR, EST and in silico methods (Bennett et al. Glycobiology 2012), which further support the use of our strategy. Finally, as will be apparent from this review as well as the data presented in the cell lines analyzed with the SimpleCell strategy in the current manuscript, most cells mainly express the more ubiquitous GalNAc-Ts (T1-T4) while other isoforms are much more restricted. We have now included additional immunocytochemistry data on two more GalNAc-Ts. We also want to point out that GalNAc-T20 as well as the subgroup T8/9/18/19 have not been shown to be active GalNAc-transferase enzymes and several papers as well as presentations at meetings suggest that they are ER proteins without GalNAc-T functions.
The strategy forward related to this comment is really differential O-glycoproteomes in isogenic cell lines without individual GalNAc-Ts as stated in the Summary and outlook. We did publish one example (Schjoldager et al PNAS 2012), but now have data on two more enzymes and it is quite amazing how the in vitro enzyme data predicts unique functions of these enzymes. Obviously this is preliminary data and cannot be presented at this stage.

Action: We have included analysis of two more GalNAc-Ts (T10 and T13) in suppl. Table S3.

Query 3. Unfortunately, 'unstructured' regions rich in the amino acids Pro, Thr and Ser have so far not got

a common name in databases like PFAM etc. Lang et al (PNAS (2007) 104, 16209) introduced the term PTS (-domain, -sequence, -region) for such sequences that become O-glycosylated and when long generate a mucin domain. The authors are encouraged to use nomenclature especially as they use the same letters, but in another order. Alter p. 7, STP, and p. 12, PST two times into PTS. Also rephrase and add PTS to the 'tandem repeat region' page 7, l. 9. Some PTS regions have VNTR, but not all.
Action: PST has been changed into PTS on page 12 and rephrased in page 7.

MINOR
Query 1. p. 5, l. 18: Alter breadth to depth.

Answer 1.
Action: changed (now Page 6, Line 15)

Query 2. p. 7-8: Should have been very helpful for the community to pull out the 14% proteins that were classified as nuclear into a separate supplementary Table. Among these are probably significant proteins taking part in signaling.

Answer 2.
While this may be an interesting exercise per se such an analysis is heavily flawed by incorrect and redundant annotations of proteins. This type of data is primarily used in the literature to reflect relative representation and rarely used alone to define localization of individual or subsets of proteins. As a matter of fact the 14% of proteins annotated to nucleus (65 in total) in this study in many cases have multiple annotations and obvious incorrect annotations. We have included an excel sheet for the reviewer listing the 65 proteins and it will be clear to the reviewer that several have signal peptides and conflicting annotations. We have now manually perused the annotation of the 65 proteins and of these most had multiple annotations and only a subset of 25 as listed here had unambiguous nuclear localisation annotations. However, even among these there are proteins with obvious signal peptides (e.g. Q6H9L7).
Q07954, Q6ZMZ3, Q6H9L7, Q9UNA0, Q9C002, O95084, Q93052, P01011, Q9UH99, Q8TEY5, Q9BVT8, Q3T906, Q96BA8, Q86VF2, O43660, Q9Y2G1, P18850, P42857, Q96A49, O60502, O14657, P98172, O94901, Q99941, O75976
We feel that it would be misleading to include specific protein lists without serious manual (and experimental) validation to guide readers to protein subsets of particular interest. We have instead disclosed detailed information on how to use our dataset to obtain these lists for anybody interested in following the method using Cytoscape as described in Fig. 4 legend.

Action: We have included the following text in Fig. 4 legend to reflect the above:
“Cellular components that are significantly over (A) or under (B) represented in the SimpleCell data set compared to the entire human proteome using BinGO plugin for Cytoscape (www.cytoscape.org). The individual annotations for each protein have not been validated manually as the analysis is merely for relative representation purposes.

Referee #3
Query 1) Why is the existence of Tyr-O-GalNAc referred to as "likely" in abstract? Were the calls ambiguous?

Answer 1.
The calls were not ambiguous and O-glycosylation of Tyr found both by us and the Halim/Larson group as discussed. The “likely” refers to whether GalNAc-Ts are involved in the biosynthesis and we believe that this is unambiguously stated as is? Furthermore, the Results section contains the following paragraph for explanation:

"One recent surprise in the field has been the identification of GalNAc O-glycosylation of Tyr residues. The first site was identified in an intriguing position in the amyloid P protein close to the β'-processing site of BACE-1 (Halim et al., 2011), and we identified sites in several other proteins (Steentoft et al., 2011). Here, we identified another 17 sites in diverse proteins, to bring the total to 23 identified Tyr O-glycosites (Table S2G). It is unclear if these sites are glycosylated by GalNAc-Ts, but in preliminary studies we have observed incorporation of GalNAc on Tyr in peptide substrates by in vitro enzyme assays (not shown)".
The Discussion contains the following sentence:
"We presume that the biosynthesis of Tyr O-glycosylation is controlled by GalNAc-Ts, but further studies are needed to evaluate this."

Action: None

Query 2) The qualifications relating to the general strategy of isolating O-GalNAc peptides should be spelled out in the first section of the Results and Discussion. For example, the addition of Gal to form the T antigen on a glycoprotein may promote or inhibit the transfer of GalNAc to a "nearby" Ser/Thr; dense regions of O-GalNAc are/may be resistant to proteolysis; O-GlcNAc on cytosolic regions or contaminating proteins or on glycoproteins in the secretory pathway with EGF repeats containing a particular consensus sequence (defined in (Alfaro et al (2012) PNAS 109, 7280)) cannot be distinguished from O-GalNAc by mass; only O-GalNAc glycopeptides that bind to VVA will be detected.

Answer 2.
We agree.

Action: The following section has been added to start of the Results (Pages 5-6):
"The results present a first generation view of the GalNAc O-glycoproteome, and we expect that analysis of additional cell lines, use of different proteases to enhance coverage, prior release of e.g. N-glycans to identify glycopeptides with both N- and O-glycosites, and use of instrumentation with enhanced sensitivity, such as the OrbiTrap Velos Pro or Elite, will lead to substantial increases in number of detected O-glycoproteins and glycosites. Furthermore, it is important to note that the SimpleCell strategy may suffer from several shortcomings: i) elimination of the O-glycan elongation pathway in cells may facilitate enhanced density of O-GalNAc glycosylation due to lack of competition with the lectin mediated functions of GalNAc-Ts (Bennett et al., 2012), although we have found no evidence of such, and a comparative analysis of the O-glycoproteomes of wild type and SimpleCells of K562 showed they were similar (Steentoft et al., 2011); ii) densely O-glycosylated regions may be less susceptible to proteolysis, and hence poorly detected by the mass spectrometric approach; however, this should be less pronounced with short GalNAc O-glycans, as evidenced by many identified O-glycopeptides with proteolytic cleavage in close proximity or adjacent to glycosites; and iii) mass spectrometry can easily identify HexNAc modifications, but can not distinguish between the isomeric/isobaric O-GalNAc and O-GlcNAc residues; however, O-GlcNAc is primarily found on cytosolic proteins without signal sequences (Hart and Akimoto, 2009) or, as recently demonstrated, in some EGF-like repeats on Notch and a few other glycoproteins (Alfaro et al., 2012; Sakaidani et al., 2011)."

Query 3) Were (there?) O-GalNAc glycopeptides that did not bind to the VVA column during LWAC analyzed by MS/MS? If so, the results should be summarized.

Answer 3.
The short answer is none detectable, but sensitivity of all assays used to analyze this may be a problem. The LWAC used were optimized with standard glycopeptide mixtures comprised of unglycosylated peptides and GalNAc-glycopeptides with one or more GalNAc residues to probe efficiency (Steentoft et al. Nature

Methods 2011), and we found complete and efficient isolation of glycopeptides regardless of number of GalNAc residues using both MS and immunoassays. Using the same metodology we cannot see any GalNAc glycopeptides in the flowthrough of VVA LWAC.

Action: none

Query 4) The O-GlcNAc question should be explicitly addressed in the main text. In Methods it is mentioned that glycopeptides in cytosolic domains are likely to represent O-GlcNAc. However, it is now clear that glycoproteins in the secretory pathway other than Notch receptors carry O-GlcNAc (Alfaro et al (2012) PNAS 109, 7280). This reference should be quoted and discussed.

Answer 4.
We originally had the following statement in the Results section:
"Since GalNAc and GlcNAc are exact isobars, it is also important to minimize the potential for inclusion of O-GlcNAc sites found on cytosolic proteins without signal peptides (Hart & Akimoto, 2009) as well as distinct O-GlcNAc glycosylation of Notch EGF repeats (Sakaidani et al., 2011)."

Action: We have further included (as mentioned above) the following in the Results section:
"; and iii) mass spectrometry can easily identify HexNAc modifications, but can not distinguish between the isomeric/isobaric O-GalNAc and O-GlcNAc residues; however, O-GlcNAc is primarily found on cytosolic proteins without signal sequences (Hart and Akimoto, 2009) or, as recently demonstrated, in some EGF-like repeats on Notch and a few other glycoproteins (Alfaro et al., 2012; Sakaidani et al., 2011).""

Query 5) Are glycopeptides from glycoproteins with no signal peptide actually cytosolic, O-GlcNAcylated proteins? They should be compared to glycopeptides previously identified in O-GlcNAc databases, especially the recent analysis by Alfaro et al (PNAS (2012) 109, 7280).

Answer 5.
Thank you for pointing this out, this is a good idea and we have now checked our (glyco)peptides against the Alfaro and Burlingame recent data and only two peptides overlapped with the O-GlcNAcylated peptides. While this confirms that these two are indeed likely O-GlcNAc glycopeptides, it is clear that we cannot conclude the reverse as neither the O-GlcNAc nor the O-GalNAc glycoproteomes are fully clarified.

Action: The two sites from a single peptide on one protein (EBP) and a single peptide on another protein (C9orf172) that had positions comparable to peptides previously identified as GlcNAc glycopeptide have been marked as potentially GlcNAc in the Supplemental Table S2.

Query 6) Validation of the superior prediction capabilities of NetOGlyc v4.0 should be spelled out by identifying O-GalNAc glycosites predicted by v4.0 but not v3.0, and actually identified by MS/MS in this or other studies.

Answer 6.
Although the validation of the predictor does illustrate the improved predictive capability of the predictor, we understand the need to be able to see changes in prediction in a more concrete way. Lists of sites identified by 4.0 and not by 3.1 are somewhat large, and will not be particularly easy to digest. Thus, we feel it is best to present this information in the framework of the GlycoDomainViewer.

Action: We have as a utility provided a comparison capability to the GlycoDomainViewer (http://glycodomain.glycocode.com/comparison/), where NetOGlyc3.1 and NetOGlyc4.0 predictions can be visualized for individual proteins.

Query 7) The description of the GlycoDomain Viewer should be clarified. For example, the following sentences in Results are confusing:
"Using the GlycoDomainViewer we determined that over 80% of the sites occurred outside the curated set of domains. In contrast, N-glycosylation motifs appeared within domains in 78% of the referenced annotated sites in Uniprot of our dataset."
---80% of which sites? The N-glycosylation motif is not defined. Were the broadest N-glycosylation motifs used according to the most recent proposals?

Answer 7.
Thank you, this text was somewhat unclear. The GalNAc O-glycosylation sites used for the analysis were based on our SimpleCell data and 80% were placed outside annotated domains. The N-glycosylation sites used for the analysis were based on UniProt annotations, but filtered to only include sites with experimental references. The text has been updated to clarify this.

Action: The text in Results section page 8 have been changed to the following:
"In order to enable more detailed studies of positions of O-glycosites in proteins and predictions of potential functions of site-specific O-glycosylation and relationship between protein structure and glycosylation, we have developed a graphic tool (GlycoDomainViewer) that incorporates curated protein domain annotations, as well as both identified and predicted sites of N- and O-glycosylation (N-glycosylation sites based on Uniprot data) (http://glycodomain.glycocode.com).
Using the GlycoDomainViewer we determined that over 80% of the sites identified by our SimpleCell strategy were located outside the curated set of domains. In contrast, analyzing the same protein subset for UniProt annotated N-glycosylation sites (only sites with experimental references included), we found that approximately 78% of the N-glycan sites were located within annotated protein domains."

Query 8) An expanded legend to Fig. 5 explaining the diagram is necessary. How were N-glycan sites identified? How were N-glyosites predicted? How were O-glycosites predicted? Better to use N rather than a chitobiose core symbol for N-glycans. Identified N-sites could be bolded N.

Answer 8.
We decided to use chitobiose as it is unique for N-linked glycans and adding a mannose would create a spacing problem in the figure. We believe a broader spectrum of people will recognize the chitobiose symbol compared to an N when the O-linked is presented as symbols.

Action: The legend has been expanded to further clarify.

Query 9) Predicted O-GlcNAc modifications if any, should be included in Fig. 5.

Answer 9.
This has been discussed also in relation to query 5., but essentially only two sites were overlapping and hence predicted to be O-GlcNAc. As far as we know there are no good predictor of GlcNAc O-glycosylation sites and the proteins used as examples for Fig. 5 have not been found experimentally to our knowledge to have GlcNAc O-glycosylation, although it is highly likely that the cytosolic domains of the receptors in fact are. Also none of these represent Notch EGF domain proteins.

Action: None

Query 10) The number of v3.1 glycosites in Fig. 6C is extremely high. Appears to be an error.

Answer 10.
Yes, thank you for pointing this out. This was a display error due to two different versions of Illustrator.
Action: Has now been fixed.

Query 11) The legend to Fig. 7 should be expanded to clarify the figure. Fig. 7 itself is hard to read. Bolding the T enzymes may help.

Answer 11.
Agree.

Action: We have altered the figure and expanded the legend to Fig. 7.

Query 12) The in vitro assay used on 181 peptides represents a very crude comparison as noted by the authors. The Discussion should briefly present how to go about this comparison to obtain better answers in a larger study.

Answer 12.
We agree.

Action: Following paragraph added to Results Section:
"This is clearly a first snapshot of isoform specificity and further studies with larger sets of peptides, combinations of enzymes to probe follow-up functions, analysis of sites utilized, and inclusion of more GalNAc-T isoforms are necessary to built datasets useful for elucidating consensus motifs for individual GalNAc-T isoforms. In this respect further studies of differences in O-glycoproteomes obtained from isogenic cells with loss (or gain) of specific GalNAc-T isoforms as recently demonstrated with GalNAc-T2 in HepG2 should be highly complementary (Schjoldager et al., 2012)."

Query 13) The manuscript should specifically address the O-glycosite/GalNAcT isoform consensus sequences previously proposed by Gherken and others in relation to the ~3000 glycosite sequences reported here. When all available data are taken into account, does a common consensus sequence emerge? Do the 181 peptide substrates fit predictions?

Answer 13.
This is an interesting question, which we have investigated, but no informative results could be extracted. Applying Gerken algorithms developed for a few specific GalNAc-T isoforms on the total dataset does not provide information, but we have now revised Fig. S3 to include a Weblogo plot of the entire dataset, where a general abundance of prolines are seen. However, as expected no clear consensus sequences are observable from the aggregate data. We also now tested Gerkens predictor.

Action: We have included a Weblogo plot of the sequences surrounding the glycosites as part of Supplemental Figure S3 to illustrate the lack of clear consensus sequence emerging from the data.
We have also included the following text in Results section p. 10:
"We analysed the 181 peptide substrates tested by in-vitro glycosylation (Figure 7 and Table S4) with Gerkens IsoGlyP predictor (Gerken et al., 2011). The positively in vitro glycosylated peptides were predicted quite correctly in 84%, 96%, 82%, 91%, 100% and 88% of the cases for GalNAc-T1, T2, T3, T5, T12 and T16 isoforms, respectively. However, the predictor predicted glycosylation of 43 - 55% of the peptides that were not glycosylated in vitro by the respective enzymes. Thus, there is relatively poor agreement and either the in vitro analysis under predicts or the IsoGlyP over predicts glycosylation."

Minor points:
a) The abbreviations used in Fig. 4 (ext. cel. -- etc.) are not intuitive or usual.
Action: Changed in figure and abbreviation explained in legend.

b) The Gill review is 2010 in bibliography -- should be 2011.
Action: Changed

c) Sentence needs changing: GalNAc-T isoforms differs in cells and tissues and changes during - delete second "and"
Agree.
Action: Sentence changed to:
"The repertoire of GalNAc-Ts in cells directs O-glycosylation, with several studies demonstrating that the expression of individual GalNAc-T isoforms differ in cells, tissues, during cell differentiation and also malignant progression"

d) The Discussion suggests that this paper introduces the SC strategy and should be modified.
Agree.
Action: Rephrased to:
"Introduction of the SimpleCell strategy (Figure 1) have for the first time enabled a proteome-wide discovery of the O-glycoproteome and determination of sites of O-glycan attachments (Steentoft et al., 2011)."

e) Table S3 needs a legend to describe mAb row and the numbers after each cell line.
Agree
Action: We have changed the legend and deleted numbers after each cell line as they represent clone number of the individual clones which are irrelevant information to the reader.

f) "GalNAc-Type O-Glycoproteome" is unwieldy in the title. Suggest "O-GalNAc Glycoproteome" as a more specific and enduring title.
Agree.
Action: title changed to:
Precision Mapping of the Human O-GalNAc Glycoproteome through SimpleCell Technology

2nd Editorial Decision 26 February 2013

Thank you for submitting your revised manuscript for our consideration. I have now had a chance to read it again and to carefully assess your answers to the original reviewers. In particular, I have considered your arguments in relation to the two requests for experimental extensions made by two referees, and I agree that these additional studies would be unlikely to offer useful additional insights within the scope of the present study. Therefore, I am happy to say that in light of the responses and the changes made, we shall now in principle be able to accept the paper for publication as a 'Resource' article in The EMBO Journal (note that the 'Resource' labeling is solely intended to point the appropriate target readership to articles that may be of methodological or dataset value to them, and to advertise the journal's interest in receiving and considering such studies).

Before we shall be able to proceed with formal acceptance of the paper, there remain some editorial issues to be addressed:

- first, regarding the GlycoDomainViewer described here for the first time and currently only accessible via a reviewers' password, it is my understanding that this will be made publicly and freely available upon publication. Is this correct, and how will this be done?

- second, on our routine pre-acceptance CrossCheck analysis of the manuscript text, I noticed that sizable passages of the introduction (bottom 1.5 paragraphs on page 2, bottom paragraph on page 3) appear to be near-verbatim copies of corresponding sections in previous publications of yours (Schjoldager & Clausen BBA 2012, Steentoft et al Nat Meth 2011). While I realize that these are sentences from your own earlier work, please understand that this may nevertheless be formally considered self-plagiarism if published in this form. To avoid such problems, I would kindly ask you to slightly alter/re-phrase these sections and email a modified manuscript text file back to me as early as possible.

Once these points will have been clarified, we should be in a position to move towards formal acceptance and production of the study!

Additional correspondence (author) 04 March 2013

Thank you very much for your help with our manuscript and we are very pleased with your decision and look forward to seeing our study published.

Regarding your queries:
1. You are correct that we intend to have access to GlycoDomainViewer freely available on or before the paper is published. In order to provide a stable service to end users when we make it publicly available, we have managed to get an agreement to move the server for the GlycoDomainViewer across from the current location (http://glycodomain.glycocode.com<http://glycodomain.glycocode.com/>) to a more permanent location, hosted somewhere under the http://cbs.dtu.dk<http://cbs.dtu.dk/> website on more powerful servers. You may know that CBS hosts a large collection of prediction servers and is widely used. The tool will be available there without passwords or restrictions, and will also have links to download the source code. We believe it will take around a week or so to co-ordinate, at which point we can both provide an updated final url for the tool, as well as the associated changes in the text in the paragraphs you indicated.

2. Thank you very much for pointing this out and we are editing the text to prevent this. This was entirely my fault and I will be more careful in the future not to fall in love with my own phrases!

If this is acceptable we can submit a revised version with the correct url and edited text within 2 weeks.

2nd Revision - authors' response 15 March 2013

Enclosed please find our revised manuscript as agreed. We have edited the two paragraphs with resemblance to our previous papers and now included the final URL address for the GlycoDomainViewer (http://cbs.dtu.dk/biotools/glycodomainviewer). The viewer should be up and active early next week without password.

Acceptance letter 18 March 2013

Thank you for sending us your final revised manuscript for our consideration. I am pleased to inform you that we have now accepted it for publication in The EMBO Journal.

Thank you again for this contribution to The EMBO Journal and congratulations on a successful publication! Please consider us again in the future for your most exciting work.

