Title:

An ultrasensitive method for analysis of viral spike N-glycoforms

Authors:

Sabyasachi Baboo*, Jolene K Diedrich, Salvador Martínez-Bartolomé, Xiaoning Wang, Torben Schiffner, Bettina Groschel, William R Schief, James C Paulson & John R Yates III*

Affiliations:

Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA, USA

Sabyasachi Baboo, Jolene K Diedrich, Salvador Martínez-Bartolomé, Xiaoning Wang, James C Paulson and John R Yates III

IAVI Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, CA, USA

Torben Schiffner, Bettina Groschel, William R Schief

Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA

Torben Schiffner, Bettina Groschel, William R Schief, James C Paulson

The Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA

Torben Schiffner, William R Schief

Present address:

Institut für Wirkstoffentwicklung, Universität Leipzig, Leipzig, Germany

Torben Schiffner

Corresponding authors:

Correspondence to Sabyasachi Baboo (sbaboo@scripps.edu) or John R Yates III (jyates@scripp.sedu)

Keywords:

proteomics, mass spectrometry, Proteinase K, N-glycan, heterogeneity, HIV, Env, viral spike

Words:

1,462 (abstract/70 + main)

Abstract:

Viruses can evade the host immune system by displaying numerous glycans on their surface "spike-proteins" that cover immune epitopes. We have developed an ultrasensitive "single pot" method to assess glycan occupancy and the extent of glycan processing from high-mannose to complex forms at each N-glycosylation site. Though aimed at characterizing glycosylation of viral spike proteins as potential vaccines, this method is applicable for analysis of site-specific glycosylation of any glycoprotein.

Main:

Viral spike-proteins initiate virus entry into host cells and are the primary targets of vaccine design. Spike-proteins are often heavily N-glycosylated which help to shield the protein from the host immune response[1]. These glycans add complexity to the production and characterization of

47 recombinant protein-based vaccines[2]. This is particularly a concern for characterization of the
48 envelope spike-protein (*Env*) of the Human Immunodeficiency Virus (HIV) comprising a trimer
49 with each monomer containing 26-30 unique N-linked glycosylation sites (NGS) as defined by
50 the sequon `NX[S|T]`, where `X` is any amino acid except `P`[3]. To address this analytical challenge
51 several mass spectrometry-based strategies using multiple proteases[4-6] have been implemented to
52 create sufficient numbers of peptides unique to each glycosylation site[3, 7, 8]. In these strategies,
53 individual aliquots are digested with each protease and analyzed separately by liquid
54 chromatography-mass spectrometry (LC-MS/MS), or pooled and analyzed together. To broadly
55 characterize the nature of the glycosylation at each NGS, we had previously introduced the use
56 of endoglycosidases that create residual mass signatures[3]. This helped us to determine the degree
57 of glycan occupancy, and the degree of glycan processing – from the high mannose form that is
58 initially attached to the protein, which may mature into the complex form when mannose
59 residues are replaced by "terminal" monosaccharide sequences. To achieve coverage for all
60 NGS, we had combined several digestions performed with different proteases and achieved
61 >95% sequence coverage. Here we show that it is possible to replace these multiple proteolytic
62 digestions with a single Proteinase K (PK) digestion, and moreover through careful choice of
63 volatile buffers we have developed an improved "single pot" strategy with significantly
64 increased sensitivity.
65
66 PK is a broadly specific serine-protease that has previously been exploited for its potential to
67 generate overlapping peptides and high sequence coverage[9]. The redundancy afforded by
68 overlapping sequences significantly increases confidence in identifications, especially when
69 covalent modifications are present[4]. However, because proteinase K is an aggressive protease it

70 is necessary to attenuate its proteolytic activity to obtain high sequence coverage of proteins.
71 Attenuation of PK can be achieved using suboptimal reaction conditions[9] (to reduce the rate of
72 enzyme activity) and limited reaction time. Using a mildly acidic, chaotrope-free solution to
73 attenuate the activity of PK, we were able to achieve >95% sequence coverage of candidate viral
74 spike proteins and identify all NGS. The proteomic strategy of our new method is conceptually
75 similar to our previous approach[3], but it is significantly faster and more sensitive. These
76 improvements result from three key changes to the strategy: [1] using only mass spectrometry-
77 compatible constituents, samples are processed in a single solution except the final step of
78 PNGase F deglycosylation; [2] reaction volumes are kept to a minimum (5-8 µl) to increase the
79 rate of reaction to limit sample loss on surfaces and minimize freeze-drying time; and [3] the use
80 of PK provides faster digestion and excellent sequence and NGS coverage with less starting
81 material. Our new method reduces sample preparation time from 3 days to 1 day, reduces LC-
82 MS/MS run times by 9-24 fold, and can achieve 50-180 times higher sensitivity than existing
83 methods[3, 7, 8]. In addition, we have developed the data analysis tool GlycoMSQuant which further
84 reduces analysis times and simplifies data analysis.
85
86 We tested our approach on BG505 SOSIP.664 MD39[10], a stabilized native-like HIV *Env* trimer
87 being developed for an HIV vaccination strategy[11] targeting germline precursors of broadly
88 neutralizing antibodies (bNAbs) that are impacted by N-glycans. As in our previous approach[3],
89 the glycosylated peptides generated by PK digestion were sequentially deglycosylated; first with
90 Endo H to remove high-mannose and hybrid N-glycans, and then with PNGase F, which
91 removes all remaining N-glycans. The resulting residual masses on asparagine (N) in NGS is
92 +203 Da at sites occupied by high-mannose/hybrid N-glycans or +3 Da at sites occupied by

93 complex N-glycans when PNGase F deglycosylation is carried out in the presence of $H_2{}^{18}O$
94 (differentiating these sites from any deamidated Ns). Unoccupied NGS results in no (+0 Da)
95 residual mass on N. Using this new method (***Fig. 1a***), we achieved >99% amino acid sequence
96 coverage and identified all theoretically possible 27 NGSs from a single LC-MS/MS run of 0.5
97 µg of peptides generated from a starting material of 5 µg purified protein (***Fig. 1b*** and
98 ***Supplementary Figures 1a,b***). We used semi-quantitative label-free analysis based on precursor
99 peak areas to calculate the proportion of N-glycan occupancy (unoccupied: complex: high-
100 mannose/hybrid N-glycans) for each NGS. We reanalyzed the N-glycan microheterogeneity
101 pattern on BG505 SOSIP.664[12] HIV *Env* trimer from data obtained using our previous approach[3]
102 and compared it with results using the method we describe here (***Supplementary Figures 2***). The
103 results with improved method were highly comparable to those using our original approach, in
104 spite of being processed differently and the samples being prepared at different times in different
105 laboratories.
106
107 Initial results demonstrated that the new approach is at least 18 times more sensitive than our
108 previous approach[3] even though it uses a simpler and shorter workflow. To evaluate the limit of
109 sensitivity of our method, we processed progressively decreasing amounts of starting material,
110 ranging from 1 µg to 5 ng. We observed that a single LC-MS/MS run with 1 µg of starting
111 material was enough to cover >95% of the amino acid sequence and all NGS (***Fig. 2a***), which is
112 90 times more sensitive than our previous approach[3]. Major differences in microheterogeneity at
113 each NGS were generally observed when we started with <100 ng material (***Supplementary***
114 ***Figure 1c***). This is likely due to low sampling as evidenced by a decrease in amino acid

115 sequence and NGS coverage (***Fig. 2a***), as well as the absolute number of identified peptides
116 representing each NGS (***Supplementary Figure 1d***).
117
118 The improved method is agnostic to mass spectrometry platform (***Fig. 2b***). A timsTOF Pro mass-
119 spectrometer coupled to an Evosep One HPLC (timsTOF/Evosep)[13] was used to achieve >99%
120 sequence coverage and identification of all NGS using a single LC-MS/MS run with 0.5 µg of
121 starting material and an 88-minute LC gradient (***Fig. 1c*** and ***Supplementary Figures 3a,b***).
122 Thus, the sensitivity of our method on this platform was 180 times higher than our previous
123 approach[3].
124
125 N-glycan heterogeneity reflects the immunogenicity of the viral spike-protein and is critical for
126 designing vaccines[14]. The reproducibility of N-glycan heterogeneity patterns obtained with our
127 new method suggests that this is a robust procedure. Variability in sequence coverage are not
128 observed until the limits of detection are reached on an LC-MS/MS platform. Although our
129 results within the same LC-MS platform were reproducible (except some variations when using
130 different proteases, ***Supplementary Figure 5***), we may infer the effects of sampling differences
131 when comparing two different LC-MS platforms (QE-HFX *vs.* timsTOF/Evosep), as in case of
132 N156, N160, N197, N386, N392 (relative peptide abundance is persistently low) and N88, N295,
133 N301, N332, N355, N406, N411 (possible skewing of timsTOF/Evosep identification against
134 N+203 peptides when peptide sampling per NGS decreases due to less starting material) (***Fig. 2b***
135 and ***Supplementary Figures 1c,d*** and ***3c,d***). When enough sampling per NGS is achieved, these
136 variations are diminished (***Figs. 1b,c***).
137

138 We attribute improvements in our method to efficient sample handling strategies. We observed
139 reduced sequence coverage if the sample was from the digestion of a small amount of starting
140 material rather than an equal aliquot from a larger sample (***Supplementary Figure 3e***). We infer
141 that the sensitivity differences are not occurring during LC-MS/MS, but that sample is being lost
142 to the reaction-tube surface (during reaction and lyophilization) and the proportion of loss is
143 more pronounced when we start with less material. The kinetics of the enzyme-substrate reaction
144 may also account for sensitivity differences since a more "crowded" reactant environment (low
145 reaction volumes) is expected to result in better reaction kinetics[15].
146
147 The simplicity and high reproducibility of this procedure will allow for high-throughput analyses
148 of viral spike-proteins and for any glycoprotein whether produced recombinantly or purified
149 from natural sources. Results were highly comparable in the two LC-MS/MS platforms we used.
150 A single analysis using a QE-HFX/nLC platform can determine the complete N-glycan
151 heterogeneity pattern from 1 microgram of purified viral spike-protein. The timsTOF/Evosep
152 platform was observed to be more sensitive than QE-HFX/nLC across all NGS, although limited
153 sampling may not allow us to confidently infer N-glycan microheterogeneity at all NGS. We
154 view the high sensitivity of this robust method to be an important step to analysis of the
155 glycosylation of more complex samples such as whole virus or virus in infected blood[16].
156
157 **Methods:**
158
159 <u>Expression and purification of HIV *Env* trimers</u>

160 BG505 SOSIP.664[12] and BG505 SOSIP.664 MD39[10] *Env* trimers were expressed and purified
161 essentially as described previously[10]. Briefly, sequences with codons optimized for expression in
162 human cells were synthesized and cloned into pHLSec between *AgeI*/*KpnI* by Genscript. The
163 constructs were co-transfected with Furin-encoding plasmid, using polyethylenimine in Freestyle
164 293F cells cultured in 293 FreeStyle media (Thermo Fisher Scientific). Where indicated, 15 µM
165 sterile-filtered Kifunensine (Cayman Chemical) was added after transfection. After 6-7 days,
166 supernatant was collected after passing through 0.22 µm filter (Nalgene), and the C-terminally
167 His-tagged trimers were purified using a HisTrap affinity column (Cytiva) with a linear elution
168 gradient from 20-500 mM imidazole, followed by a Superdex 200 Increase SEC column (Cytiva)
169 in Tris-buffered saline/TBS (20 mM Tris, 100 mM NaCl, pH 7.5). The oligomeric state and
170 purity of trimer was verified using size exclusion chromatography coupled with multi-angle light
171 scattering (SEC-MALS; DAWN HELEOS II/ Optilab T-rEX, Wyatt Technology).
172
173 Proteinase K treatment and deglycosylation
174 HIV Env trimer was exchanged to water using Microcon Ultracel-10 centrifugal device
175 (Millipore Sigma). Trimer was reduced with 5 mM tris(2-carboxyethyl)phosphine hydrochloride
176 (TCEP-HCl, Thermo Scientific) and alkylated with 10 mM 2-chloroacetamide (Sigma Aldrich)
177 in 100 mM ammonium acetate for 20 min at room temperature (RT, 24°C). Initial protein-level
178 deglycosylation was performed using 250 U of Endo H (New England Biolabs) for up to 5 µg
179 trimer, for 1 h at 37°C (pH 5.5-6.0). Trimer was digested with 1:25 Proteinase K (Sigma
180 Aldrich) for 4 h at 37°C (pH 5.5-6.0). PK was denatured by incubating at 90°C for 15 min, then
181 cooled to RT. Peptides were deglycosylated again with 250 U Endo H for 1 h at 37°C (pH 5.5-
182 6.0), then frozen at –80°C and lyophilized. 100 U PNGase F (New England Biolabs) was

183 lyophilized (for up to 5 µg trimer), resuspended in 100 mM ammonium bicarbonate prepared in
184 $H_2{}^{18}O$ (97% ^{18}O, Sigma-Aldrich), and added to the lyophilized peptides. The resulting 5-8 µl
185 reaction solutions (except PNGase F reaction, in 5-20 µl) were then incubated for 1 h at 37°C
186 (pH 8.0-8.5) in 0.2 ml PCR tubes on a thermocycler with heated lid.
187
188 Validating efficiency of glycosidases
189 BG505 SOSIP.664 HIV *Env* trimer glycosylated with only high-mannose N-glycans (purified
190 from cells treated with Kifunensine[17], which inhibits processing of high-mannose N-glycans to
191 complex N-glycans during protein maturation), was sequentially treated with Endo H, followed
192 by PNGase F. After treatment with both enzymes, 99.2% of identified peptides were N+203;
193 100% of peptides identified with only PNGase F treatment were N+3 (***Supplementary Figures***
194 ***4a,b***; proportions do not consider unoccupied NGS because they remained similar in both
195 experiments – 8-10%). We realize the possibility that glycosidase PNGase F may occasionally
196 cleave the remnant GlcNAc (N-Acetylglucosamine)[18] post-Endo H processing of high-
197 mannose/hybrid N-glycans and thus, convert the mass modification characteristic of high-
198 mannose/hybrid (+203) to complex (+3) N-glycans, which would affect our analyses. However,
199 we have not observed any significant evidence of this possibility in our results, though it may
200 explain why we observe a few peptides with +3 mass modified NGS in Kifunensine treated
201 samples (***Supplementary Figure 4a***).
202
203 Trypsin proteolysis
204 The Proteinase K/deglycosylation method described above was followed, except PK was
205 replaced with trypsin and reactions were incubated overnight at 37°C. Trypsin generated a lower

206 total number of peptides than PK, but we obtained >95% sequence coverage, including 26 of 27
207 NGS (***Supplementary Figure 5a***). Variations in N-glycan microheterogeneity at certain NGS
208 may be explained by low sampling at these sites (N386, N392) or difference in cleavage-
209 specificity between PK and trypsin (N88, N611) (***Supplementary Figures 5b,c***).
210
211 LC-MS/MS
212 *Q Exactive HF-X with EASY-nLC 1200*
213 Samples were analyzed on an Q Exactive HF-X mass spectrometer (Thermo). Samples were
214 injected directly onto a 25 cm, 100 µm ID column packed with BEH 1.7 µm C18 resin (Waters).
215 Samples were separated at a flow rate of 300 nL/min on an EASY-nLC 1200 (Thermo). Buffers
216 A and B were 0.1% formic acid in 5% and 80% acetonitrile, respectively. The following
217 gradient was used: 1–25% B over 160 min, an increase to 40% B over 40 min, an increase to
218 90% B over another 10 min and 30 min at 90% B for a total run time of 240 min. Column was
219 re-equilibrated with solution A prior to the injection of sample. Peptides were eluted from the tip
220 of the column and nanosprayed directly into the mass spectrometer by application of 2.8 kV at
221 the back of the column. The mass spectrometer was operated in a data dependent mode. Full
222 MS1 scans were collected in the Orbitrap at 120,000 resolution. The ten most abundant ions per
223 scan were selected for HCD MS/MS at 25 NCE. Dynamic exclusion was enabled with exclusion
224 duration of 10 s and singly charged ions were excluded.
225 *timsTOF Pro with Evosep One*
226 Samples were loaded onto EvoTips following manufacturer protocol. The samples were run on
227 an Evosep One (Evosep) coupled to a timsTOF Pro (Bruker Daltonics). Samples were separated
228 on a 15 cm × 150 µm ID column with BEH 1.7 µm C18 beads (Waters) and integrated tip pulled

229 in-house using either the 30 SPD or 15 SPD methods. Mobile phases A and B were 0.1% formic
230 acid in water and 0.1% formic acid in acetonitrile, respectively. MS data was acquired in PASEF
231 mode with 1 MS1 survey TIMS-MS and 10 PASEF MS/MS scans acquired per 1.1 s acquisition
232 cycle. Ion accumulation and ramp time in the dual TIMS analyzer was set to 100 ms each and we
233 analyzed the ion mobility range from $1/K_0$ = 0.6 Vs cm^{-2} to 1.6 Vs cm^{-2}. Precursor ions for
234 MS/MS analysis were isolated with a 2 Th window for m/z < 700 and 3 Th for m/z >700 with a
235 total m/z range of 100-1700. The collision energy was lowered linearly as a function of
236 increasing mobility starting from 59 eV at $1/K_0$ = 1.6 VS cm^{-2} to 20 eV at $1/K_0$ = 0.6 Vs cm^{-2}.
237 Singly charged precursor ions were excluded with a polygon filter, precursors for MS/MS were
238 picked at an intensity threshold of 2,500, target value of 20,000 and with an active exclusion of
239 24 s.
240
241 Data Processing
242 Protein and peptide identification were done with Integrated Proteomics Pipeline (IP2, Bruker
243 Scientific LLC). Tandem mass spectra were extracted from raw files using RawConverter[19]
244 (timstofCoverter for timsTOF Pro data) and searched with ProLuCID[20] against a database
245 comprising UniProt reviewed (Swiss-Prot) proteome for *Homo sapiens* (UP000005640), UniProt
246 amino acid sequences for Endo H (P04067), PNGase F (Q9XBM8), and Proteinase K (P06873),
247 amino acid sequences for BG505 SOSIP.664[12] and BG505 SOSIP.664 MD39[10] (including a
248 preceding secretory signal sequence and followed by 6xHis-tag), and a list of general protein
249 contaminants. The search space included no cleavage-specificity (all fully tryptic and semi-
250 tryptic peptide candidates for trypsin treatment). Carbamidomethylation (+57.02146 C) was
251 considered a static modification. Deamidation in presence of $H_2{}^{18}O$ (+2.988261 N), GlcNAc

252 (+203.079373 N), oxidation (+15.994915 M) and N-terminal pyroglutamate formation (–
253 17.026549 Q) were considered differential modifications. Data was searched with 50 ppm
254 precursor ion tolerance and 50 ppm fragment ion tolerance. Identified proteins were filtered
255 using DTASelect2[21] and utilizing a target-decoy database search strategy to limit the false
256 discovery rate to 1%, at the spectrum level[22]. A minimum of 1 peptide per protein and no tryptic
257 end (or 1 tryptic end when treated with trypsin) per peptide were required and precursor delta
258 mass cut-off was fixed at 10 ppm for data acquired with Q Exactive HF-X or 20 ppm for data
259 acquired with timsTOF Pro. Statistical models for peptide mass modification (modstat) were
260 applied (trypstat was additionally applied for trypsin-treated samples). Census2[23] label-free
261 analysis was performed based on the precursor peak area, with a 10 ppm precursor mass
262 tolerance and 0.1 min retention time tolerance. "Match between runs" was used to find missing
263 peptides between runs for Q Exactive HF-X data (for timsTOF Pro data, reconstructed-MS1
264 based chromatograms combining isotope peaks for all triggered precursor ions were pre-
265 generated, and then chromatograms were assigned to identified peptides for quantitative analysis,
266 without retrieving missing peptides).
267
268 Data Analysis using GlycoMSQuant
269
270 Our new tool GlycoMSQuant v.1.4.1 (https://github.com/proteomicsyates/GlycoMSQuant) was
271 implemented to automate the analysis and to visualize the results. GlycoMSQuant summed
272 precursor peak areas across replicates, discarded peptides without NGS, discarded misidentified
273 peptides when N-glycan remnant-mass modifications were localized to non-NGS asparagines

274 and corrected/fixed N-glycan mislocalization where appropriate. The results were aligned to

275 NGS in *Env* of HXB2[24] HIV-1 variant.

276

277 Precursor peak area was calculated by Census2[23] from extracted-ion chromatogram (XIC) for

278 each peptide in each replicate. For each NGS (`NX[S|T]`, where `X` is any amino acid except `P`), the

279 "N-glycosylation state" represented by proportions for unoccupied (+0, *u*), complex (+2.988261,

280 *c*) and, high-mannose/hybrid (+203.079373, *h*) N-glycans was calculated as follows.

281

282 The sum of the precursor peak areas $S_{g,pepz}$ was calculated as:

283 $$S_{g,pepz} = \sum xic_{pepz}$$

284 where N-glycosylated peptides with the same sequences and charge were grouped together

285 (*pepz*), *g* is the N-glycosylation state ∈ *G* (*u*, *c*, *h*), and *xic* is the precursor peak area.

286 For each group (*pepz*), the abundance proportion $\%_{g,pepz}$ of each N-glycosylation state $g \in G$ was

287 calculated as:

288 $$\%_{g,pepz} = \frac{S_{g,pepz}}{\sum_{i \in G} S_{i,pepz}}$$

289 Finally, as each NGS may be covered by multiple groups (*pepz*), the proportion of each N-

290 glycosylation state *g* for a particular NGS (*ngs*) is calculated as the mean of all proportions

291 $\%_{g,pepz}$ of all groups (*pepz*) covering this NGS:

292 $$\%_{ngs,g} = \frac{1}{n_{ngs}} \sum \%_{g,pepz}$$

293 where n_{ngs} is the number of groups (*pepz*) covering a particular NGS.

294 The standard error of mean of the proportion of each N-glycosylation state $g \in G$ for a particular

295 NGS ($SEM_{ngs,g}$) was calculated as:

296 $$SEM_{ngs,g} = \frac{s_{ngs,g}}{\sqrt{n_{ngs}}}$$

297 where, $s_{ngs,g}$ is the standard deviation of $\%_{g,pepz}$ from all groups (*pepz*) covering a particular NGS.

298

299 Pairwise statistical comparisons of experiments (*a* and *b*) were performed for each $g \in G$ at each

300 NGS using proportion values $\%_{g,pepz}$ of groups (*pepz*) sharing the NGS, applying the Mann-

301 Whitney U test[25]. Testing $\%_{g,pepz,a}$ *vs.* $\%_{g,pepz,b}$ individually for *u*, *c* and *h* at each NGS, we

302 calculated *p*-values that were subjected to multiple hypothesis correction using the Benjamini-

303 Hochberg (BH) method[26]. If the corrected *p*-value was <0.05 (for *u*, *c* or *h* at any NGS), then the

304 difference was considered statistically significant.

305

306 For ***Supplementary Figure 2***, published data from our previous approach[3] was reanalyzed using

307 the data analysis workflow described here. Briefly, the data analyzed is from 3 replicates of 3

308 conditions, each separately analyzed by LC-MS/MS, with total starting material of 90 µg protein,

309 and Census2[23] label-free analysis was performed simultaneously on all 9 experiments without

310 "match between runs", and the results analyzed by GlycoMSQuant. This was compared with

311 data obtained from a single LC-MS/MS run (QE-HFX/nLC) with 0.5 µg of peptides generated

312 from a starting material of 5 µg purified protein.

313

314 **Acknowledgements:**

315 We thank Bruker Daltonics for providing access to timsTOF Pro MS. We thank Bruker

316 Scientific LLC and Robin S Park for making IP2 accessible for data analysis. We thank Titus

317 Jung in J.R.Y. lab for IT support. We thank members of the J.R.Y. lab for discussion and Claire

318 Delahunty for critically perusing the manuscript. We thank Saman Eskandarzadeh, Michael

Kubitz, and Erik Georgeson in W.R.S. lab for assistance in protein production. This work was supported by the grants P41GM103533 (NIH) and UM1AI100663, UM1AI144462, R01/R56AI113867 (NIH/NIAID). We thank Bill and Melinda Gates Foundation (BMGF) Collaboration for AIDS Vaccine Discovery (CAVD) funding to IAVI Neutralizing Antibody Center (NAC) and Ragon Institute post-doctoral fellowship to T.S.

Author Contributions:

S.B. conceived the method. S.B., J.K.D. and X.W. designed the experiments. S.B. prepared the samples and analyzed data. J.K.D. performed LC-MS/MS. S.M.B. created the GlycoMSQuant tool. T.S. and B.G. expressed and purified HIV *Env* trimers. W.R.S., J.C.P. and J.R.Y. supervised the project. S.B. wrote the paper with contribution from all authors.

Competing Interests:

The authors declare no competing interests.

Data Availability:

Mass spectrometry data has been deposited in MassIVE-KB repository and is also accessible through ProteomeXchange Consortium with identifiers MSV000087414 and PXD025990, respectively.

Code Availability:

GlycoMSQuant source code is freely available at https://github.com/proteomicsyates/GlycoMSQuant under a permissive Apache License 2.0.

342

343 **References:**

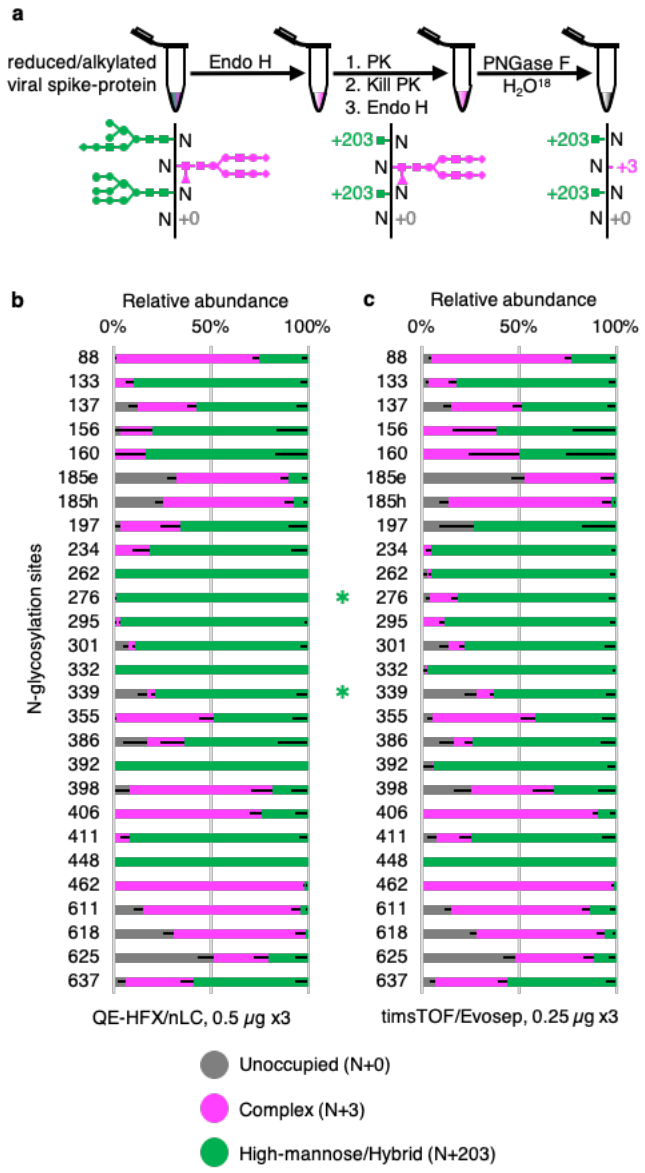
344

345 1. Rudd, P.M. & Dwek, R.A. Glycosylation: heterogeneity and the 3D structure of proteins.
346 *Crit Rev Biochem Mol Biol* **32**, 1-100 (1997).
347 2. Burton, D.R. & Hangartner, L. Broadly Neutralizing Antibodies to HIV and Their Role in
348 Vaccine Design. *Annu Rev Immunol* **34**, 635-659 (2016).
349 3. Cao, L. et al. Global site-specific N-glycosylation analysis of HIV envelope glycoprotein.
350 *Nat Commun* **8**, 14954 (2017).
351 4. MacCoss, M.J. et al. Shotgun identification of protein modifications from protein
352 complexes and lens tissue. *Proc Natl Acad Sci U S A* **99**, 7900-7905 (2002).
353 5. Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C. & Yates, J.R., 3rd Protein analysis by
354 shotgun/bottom-up proteomics. *Chem Rev* **113**, 2343-2394 (2013).
355 6. Tsiatsiani, L. & Heck, A.J. Proteomics beyond trypsin. *Febs j* **282**, 2612-2626 (2015).
356 7. Behrens, A.J. et al. Molecular Architecture of the Cleavage-Dependent Mannose Patch
357 on a Soluble HIV-1 Envelope Glycoprotein Trimer. *J Virol* **91** (2017).
358 8. Go, E.P. et al. Glycosylation Benchmark Profile for HIV-1 Envelope Glycoprotein
359 Production Based on Eleven Env Trimers. *J Virol* **91** (2017).
360 9. Wu, C.C., MacCoss, M.J., Howell, K.E. & Yates, J.R., 3rd A method for the
361 comprehensive proteomic analysis of membrane proteins. *Nat Biotechnol* **21**, 532-538
362 (2003).
363 10. Steichen, J.M. et al. HIV Vaccine Design to Target Germline Precursors of Glycan-
364 Dependent Broadly Neutralizing Antibodies. *Immunity* **45**, 483-496 (2016).

365 11. Steichen, J.M. et al. A generalized HIV vaccine design strategy for priming of broadly
366 neutralizing antibody responses. *Science* **366** (2019).
367 12. Sanders, R.W. et al. A next-generation cleaved, soluble HIV-1 Env trimer, BG505
368 SOSIP.664 gp140, expresses multiple epitopes for broadly neutralizing but not non-
369 neutralizing antibodies. *PLoS Pathog* **9**, e1003618 (2013).
370 13. Meier, F. et al. Online Parallel Accumulation-Serial Fragmentation (PASEF) with a
371 Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics* **17**, 2534-2545
372 (2018).
373 14. Seabright, G.E., Doores, K.J., Burton, D.R. & Crispin, M. Protein and Glycan Mimicry in
374 HIV Vaccine Design. *J Mol Biol* **431**, 2223-2247 (2019).
375 15. Zhou, H.X., Rivas, G. & Minton, A.P. Macromolecular crowding and confinement:
376 biochemical, biophysical, and potential physiological consequences. *Annu Rev Biophys*
377 **37**, 375-397 (2008).
378 16. Bar-On, Y.M., Flamholz, A., Phillips, R. & Milo, R. SARS-CoV-2 (COVID-19) by the
379 numbers. *Elife* **9** (2020).
380 17. Scanlan, C.N. et al. Inhibition of mammalian glycan biosynthesis produces non-self
381 antigens for a broadly neutralising, HIV-1 specific antibody. *J Mol Biol* **372**, 16-22
382 (2007).
383 18. Fan, J.Q. & Lee, Y.C. Detailed studies on substrate structure requirements of
384 glycoamidases A and F. *J Biol Chem* **272**, 27058-27064 (1997).
385 19. He, L., Diedrich, J., Chu, Y.Y. & Yates, J.R., 3rd Extracting Accurate Precursor
386 Information for Tandem Mass Spectra by RawConverter. *Anal Chem* **87**, 11361-11367
387 (2015).

388 20. Xu, T. et al. ProLuCID: An improved SEQUEST-like algorithm with enhanced
389 sensitivity and specificity. *J Proteomics* **129**, 16-24 (2015).
390 21. Tabb, D.L., McDonald, W.H. & Yates, J.R., 3rd DTASelect and Contrast: tools for
391 assembling and comparing protein identifications from shotgun proteomics. *J Proteome*
392 *Res* **1**, 21-26 (2002).
393 22. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. & Gygi, S.P. Evaluation of
394 multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-
395 MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2**, 43-50
396 (2003).
397 23. Park, S.K., Venable, J.D., Xu, T. & Yates, J.R., 3rd A quantitative analysis software tool
398 for mass spectrometry-based proteomics. *Nat Methods* **5**, 319-322 (2008).
399 24. Zhang, M. et al. Tracking global patterns of N-linked glycosylation site variation in
400 highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza
401 hemagglutinin. *Glycobiology* **14**, 1229-1246 (2004).
402 25. Mann, H.B. & Whitney, D.R. On a Test of Whether one of Two Random Variables is
403 Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18**, 50-60
404 (1947).
405 26. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
406 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B*
407 *(Methodological)* **57**, 289-300 (1995).
408
409
410

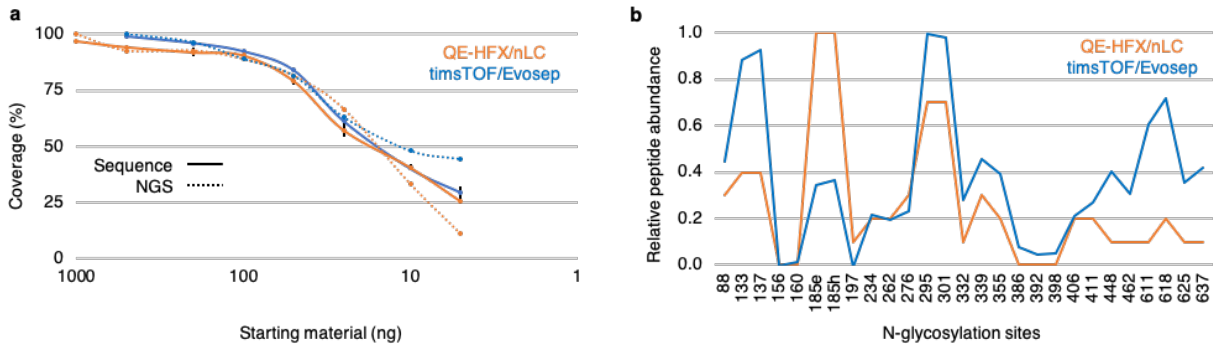
411 **Figures**



412

413 **Figure 1 | Proteinase K and glycosidase treatment provides N-glycan microheterogeneity in BG505 SOSIP.664 MD39 trimer.** (**a**) Method
414 workflow, resulting in peptides with expected mass modifications at NGS representing N-glycosylation states. (**b**) Pattern observed using 4 h
415 triplicates on an QE-HFX/nLC platform. (**c**) Pattern observed using 88 min triplicates on timsTOF/Evosep platform. N-glycosylation states are color-
416 coded. Error bars represent mean-SEM. Between (**b**) and (**c**), the pattern is similar, and any significant difference (BH-corrected *p*-value <0.05) in proportion of a certain N-glycosylation state at an NGS is represented by color-coded ∗.

417



418

419

Figure 2 I Factors affecting limit of sensitivity. (**a**) Decreasing trend of sequence and NGS coverage observed as starting material is diluted 200-fold (1 μg to 5 ng, QE-HFX/nLC) or 100-fold (0.5 μg to 5 ng, timsTOF/Evosep), using triplicates for each amount of starting material. The limit of sensitivity is revealed. (**b**) Relative abundance of peptides identified per NGS across the dilution series for the 2 LC-MS/MS platforms used. This comparison reveals a non-uniform digestion pattern that may be attributed to steric hindrance offered by the glycoprotein or characteristic behavior of individual peptides in LC-MS/MS. Error bars represent mean±SEM. Values for sequence and NGS coverage are mean

420

421

422