

Full-Length Characterization of Proteins in Human Populations

Chad R. Borges,¹ Doug S. Rehder,¹ Jason W. Jarvis,¹ Mathew R. Schaab,¹ Paul E. Oran,¹ and Randall W. Nelson^{1*}

BACKGROUND: Diversity in human proteins often gives rise to pluralities of structurally similar but functionally distinct proteins. Such microheterogeneity generally escapes proteomics discovery technologies, as well as conventional immunometric assays. As an intermediate between these 2 technological approaches, targeted, full-length characterization of proteins using mass spectrometry is a suitable means of defining microheterogeneity evident in human populations.

CONTENT: We describe and explore the implications of microheterogeneity using the exemplar of human vitamin D binding protein (Gc-Globulin) as observed in cohorts of 400 individuals. Our investigations yielded: (a) population frequency data comparable to genotyping; (b) population frequency data for protein variants, with and without genotype linkage; (c) reference values for the different protein variants per cohort and genotype; and (d) associations between variant, frequency, relative abundance, and diseases.

SUMMARY: With the exception of the genotype frequency, such population data are unique and illustrate a need to more fully understand the exact full-length qualitative and quantitative idiosyncrasies of individual proteins in relation to health and disease as part of the standardized biomarker development and clinical proteomic investigation of human proteins.

© 2009 American Association for Clinical Chemistry

Recently, a large contingent of proteomics researchers has focused their attention on defining the protein constituents of human biological fluids in health and disease. The expectation is that proteins can serve as markers of human disease, either by their presence

(disease-specific proteins) or through quantitative differences in concentrations between healthy and disease populations. Such biomarker discovery efforts often hinge on the use of proteomics approaches, in which mass spectrometry and database search engines are used to identify peptide surrogates purposely created through enzymatic proteolysis of precursor proteins. Although a rapid means of unearthing candidate biomarkers, the biomarker development process is far from complete.

Essentially, polymorphisms, transcript variants, and posttranslational modifications extend the diversity of human gene products dramatically beyond the 20 000–25 000 genes in the human genome. These multiple modifications lead to protein microheterogeneity in the human population, often giving rise to pluralities of structurally similar but functionally distinct proteins. It is only when the intact variants are separately isolated and characterized with total sequence coverage that this diversity is evident and can be correctly assigned to different products from the same gene.

Work over the past 2 decades has led to the development of high-throughput mass spectrometric immunoassay (MSIA)² technologies and workflows that are well suited to the standardized characterization of protein diversity in human populations (1–14). Data from such “population proteomics” studies include changes in concentration, amino acid sequence, and relative qualitative/quantitative changes in posttranslational modifications (15, 16). In previous work, we characterized 25 plasma proteins within a cohort of 96 healthy individuals. We observed 76 protein variants, on average 3 distinct forms of each protein evident among individuals. These variations, consisting of point mutations or posttranslational modifications, were wide ranged: we observed some modifications in only 1 sample and others in all 96 samples (15). On

¹ The Biodesign Institute, Arizona State University, Tempe, AZ.

* Address correspondence to this author at: Molecular Biomarkers, The Biodesign Institute at Arizona State University, P.O. Box 876601, Tempe, AZ 85287. Fax 480-727-9464; e-mail randal.nelson@asu.edu.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding agencies.

Received July 31, 2009; accepted October 19, 2009.

Previously published online at DOI: 10.1373/clinchem.2009.134858

² Nonstandard abbreviations: MSIA, mass spectrometric immunoassay; GcG, vitamin D binding protein (Gc-globulin); Glyc, glycosylated GcG; GalNAc, O-linked glycosylated GcG; TS, NeuNAc₁Gal₁GalNAc₁ trisaccharide; T2DM, type 2 diabetes mellitus; CVD, cardiovascular disease; GcMAF, GcG macrophage-activating factor; OEM, original equipment manufacturer; hPDQ, human proteome detection and quantification project.

expanding the studies to 1000 individuals, we observed 27 variants for 5 of the proteins, versus only 19 in the prior study (16). These results corroborated our initial findings, increased their statistical significance, and confirmed that microheterogeneity is a central and expansive feature of proteins as surveyed in large populations.

Here we illustrate the value of accurately characterizing full-length proteins in populations as part of the standardized proteomic investigation of human proteins. We center our discussions on products from a single gene [vitamin D binding protein, also known as Gc-globulin (GcG); UniProtKB/Swiss-Prot acc. no. P02774], which is generally described as a single protein, but is evident in the human population in several different forms. The accurate characterization and mapping of these variants in moderate-to-large populations finds value by defining basal-level protein differences that may confound conventional proteomics or clinical analyses or, conversely, by revealing low-level variants correlated with disease. These observations place emphasis on understanding population-based biological diversity as part of biomarker development and clinical proteomics workflows, and go toward forming an expandable “protein diversity database” that reports basal-level and determinate variants occurring in health and disease.

Unbiased, Biased, and Targeted Proteomics

For discussion, it is important to acknowledge that “unbiased” proteomics analyses generally begin with the anticipation of identifying any protein (and preferably many) as present in a sample. In this context, the term “protein” is often ascribed to an accession number resulting from searching a genomic database with mass spectral data produced from surrogate peptides. Thus, the completeness of the surrogate data, the extent of database annotation, and the fidelity of the searching algorithm are all components of this definition. In common usage, this process actually describes products resulting from particular genes with, as recently reported, variable degrees of success dependent on yet another component: the experience of the practitioner (17, 18). Quantification, as most recently practiced in proteomics, relies on selectively monitoring 1 or a few surrogate peptides derived from each gene, generally using chromatography and tandem mass spectrometry workflows akin to those used to quantify small molecules and metabolites (19–22). In proteomics terms, these investigations are now considered “biased” because they are concerned with 1 or a select group of predefined gene products. By defining these surrogate peptides a priori (and the reasons that they are analyzed), intra- and interlaboratory reproducibil-

ity are markedly improved, illustrating technical advantages stemming from the use of such biased approaches in biomarker development workflows (23, 24).

As commented recently by Duncan et al. (25), however, an *in vivo* protein may possess additional structural and semiquantitative idiosyncrasies that are not readily captured using surrogate identification and quantification approaches. By way of example, surrogate proteomics data from human blood plasma describe more than 530 sequence entries for GcG, which because of spectral redundancy align with 28 tryptic fragments corresponding to peptides stemming from genomic sequence (26). Mapping the data against the GcG sequence results in approximately 70% sequence coverage, which is commendable during the tryptic mapping of any moderate-to-large protein. Regarding quantification, a recent report describes an approach based on monitoring a surrogate peptide encompassing only approximately 2% of the GcG sequence (residues 338–347), which was found to be attenuated in approximately half the samples investigated, resulting in a technical variability of approximately 60% (24).

Important to the discussion here is not so much the extent of sequence coverage, but the gaps in the sequence (see Supplemental Fig. 1, which accompanies the online version of this article at <http://www.clinchem.org/content/vol56/issue2>). In particular, amino acids in the regions around residues 329 and 416–420 are absent. A review of the literature reveals several biological roles for GcG that include vitamin D metabolite transport, fatty acid transport, actin sequestration, and macrophage activation (27). Variable modified forms of GcG are known, as well as potential sources of such juxtaposed function and/or indication. Foremost, GcG is known to be present in human populations as 3 high-frequency allelic variants designated Gc1F, Gc1S, and Gc2. Relative to the Gc1F protein, a D416E mutation defines the Gc1S form, and a T420K mutation defines Gc2. Differentiation of these genetic forms is typically performed using conventional approaches (28, 29) but is also possible using mass spectrometry by switching to a biased analysis that automatically assumes GcG is present and then asks a further question regarding genotype. An approach is to target GcG via immunoaffinity capture and then thoroughly characterize the extracted species using workflows centering on mass spectrometry. The use of pan-antibodies, to extract multiple variants, and high-performance mass spectrometry realizes an approach that is “biased, while accounting for variability evident in the target proteins” (8, 15, 16).

Fig. 1 shows spectra resulting from MSIA analysis of GcG from 3 individuals. The methods and analytical details for the MSIA analysis have been described else-

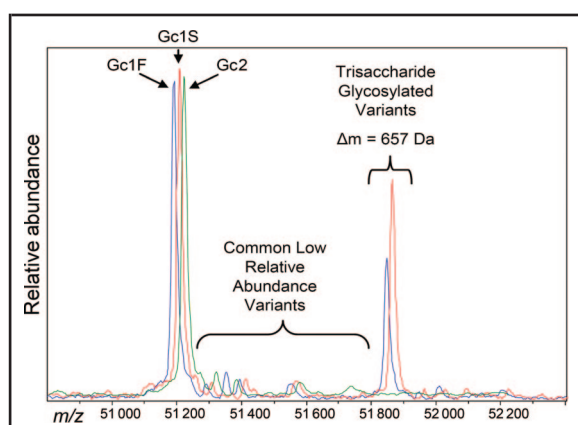


Fig. 1. Overlays of mass spectra resulting from the targeted top-down analysis of GcG from 3 individuals. Indicated are signals from the 3 major allele products Gc1F, Gc1S, and Gc2.

Also observed is native glycosylation at Δm +657 Da from their respective unmodified allele products (with the exception of Gc2). Δm , delta mass or mass shift of.

where (30, 31). Signals are observed for the 3 major genomic alleles Gc1F ($M_{\text{r,calc}}$ 51 188.2), Gc1S ($M_{\text{r,calc}}$ 51 202.2), and Gc2 ($M_{\text{r,calc}}$ 51 215.3 Da). All determined masses are within 2 Da of the calculated values. Also observed at higher molecular weight are signals

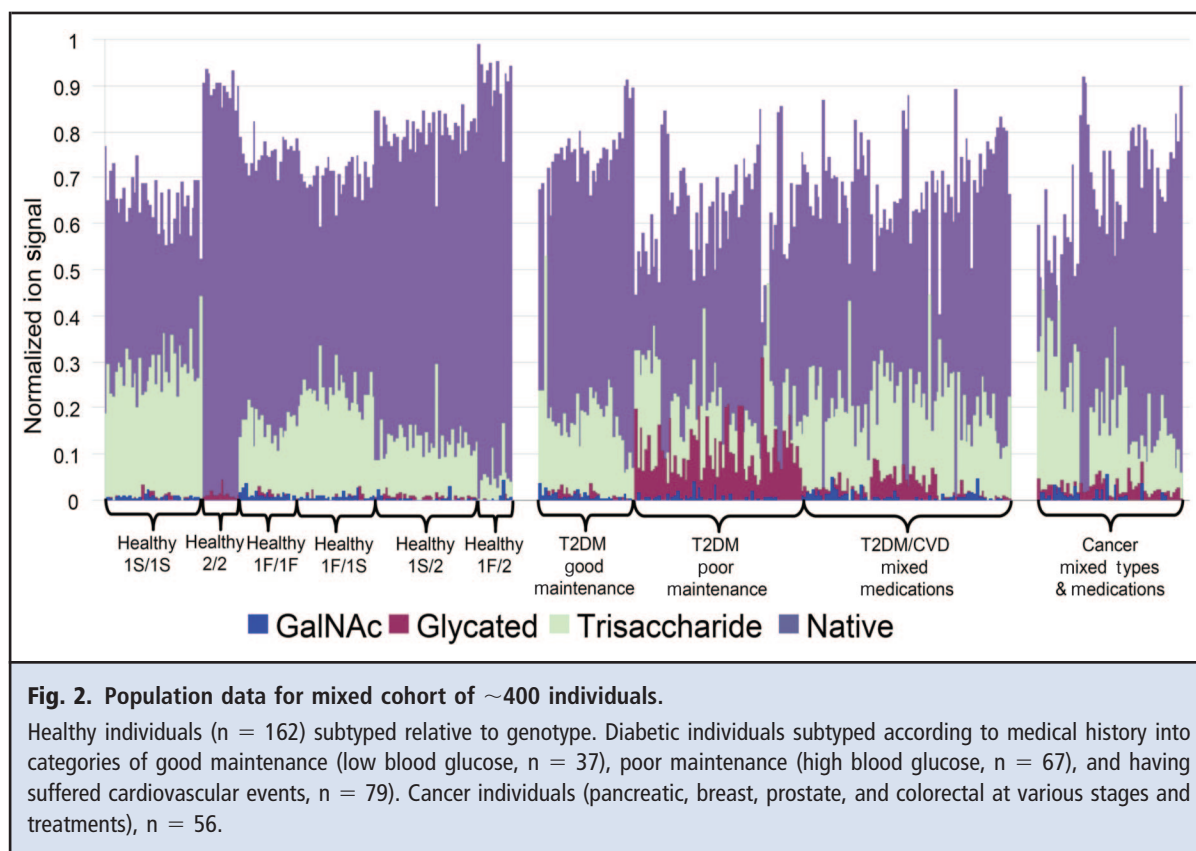
suggesting multiple posttranslational modifications stemming from the backbone precursors. Of particular interest to later discussion are signals shifted up in mass by 162, 203, and 657 Da, suggesting glycation (Glyc), O-linked glycosylation (GalNAc), and a second O-linked glycoform (NeuNAc₁Gal₁GalNAc₁ trisaccharide) (TS), respectively. Using analytical workflows that include enzymatic digestion and tandem mass spectrometry, as described (30), these identities have been confirmed and localized to glycation at site Lys³²⁹ and glycosylation in the region of residues 416–420 (see online Supplemental Fig. 2) (30). Notably, each site (or region) of interest can be responsible for at least 2 molecular species, which for thoroughness would need to be detected and monitored during surrogate searching and quantification. At residue Lys³²⁹, 2 species are possible, 1 for unmodified Lys³²⁹ and 1 for glycosylated Lys³²⁹. The region spanning residues 416–420 contains heterogeneity due to the allelic variants at residues 416 and 420 and heterogeneous glycoforms, possibly resulting in up to 12 different molecular forms of the protein. Geometrically, variability in these 2 regions of the protein can result in up to 24 different molecular variants residing in the population [i.e., unmodified, mono-, di-, and trisaccharide variants of different allelic variants that are either naked or glycosylated at Lys³²⁹ ($4 \times 3 \times 2$)], although any given individual can have up to 8 or 16 of these variants depending on homozygous or heterozygous genotype, respectively.

Table 1. GcG allelic variants and posttranslationally modified forms observed during the course of analyzing ~400 individuals.

	MH ⁺ (Δm) ^{a,b}	Identity, confirmed (putative)
Unmodified Gc1F	51 189	Unmodified Gc1F
Unmodified Gc1S	51 203	Unmodified Gc1S
Unmodified Gc2	51 216	Unmodified Gc2
Unmodified allelic variant 1	51 119	Unknown allele
PTM 1	(+162.1)	Glycation at Lys ³²⁹
PTM 2	(+203.2)	GalNAc monosaccharide
PTM 3	(+365.5)	Gal ₁ GalNAc ₁ disaccharide
PTM 4	(+656.9)	NeuNAc ₁ Gal ₁ GalNAc ₁ trisaccharide
PTM 5	(+948.0)	(NeuNAc ₂ Gal ₁ GalNAc ₁ tetrasaccharide)
PTM 6	(+266.1)	(Stearoylation)
PTM 7	(−399.2)	(des N-terminal LER)
PTM 8	(−226.3)	(des C-terminal IL or des N-terminal L and C-terminal L)
PTM 9	(−113.2)	(des N-terminal L or C-term L)

^a MH⁺ masses listed are the mean (across the population) observed average (isotopically) masses, which were within 2 Da of the predicted values (within 1 Da for Δm values).

^b MH⁺, singly protonated molecule; Δm , delta mass or mass shift of; PTM, posttranslational modification; LER, Leu-Glu-Arg (amino acid sequence); IL, Ile-Leu (amino acid sequence); L, Leu (amino acid).



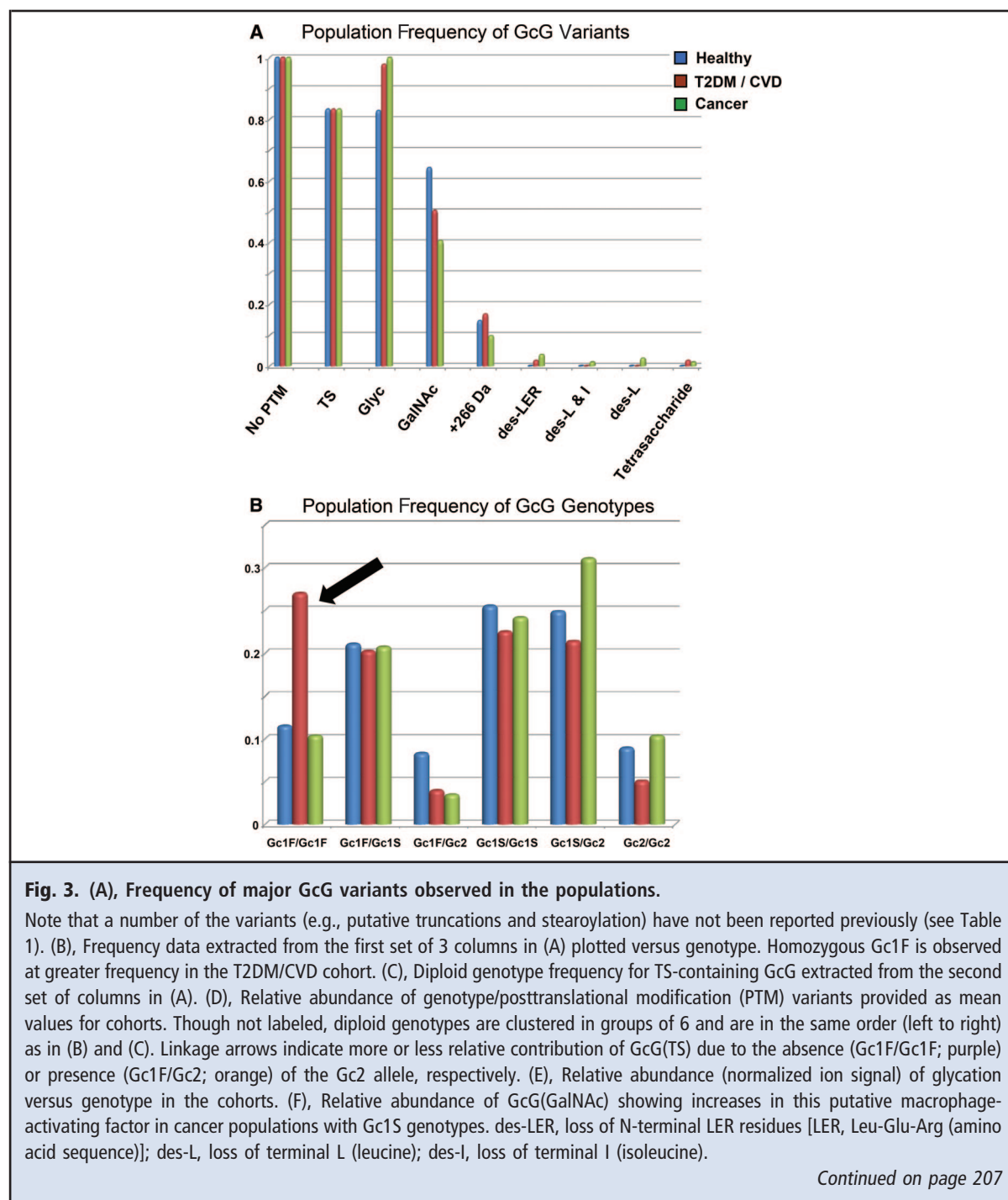
This root diversity is compounded even further when other variants are considered (Table 1), and for simplicity, in the next sections we discuss primarily the unmodified, Glyc, GalNAc, and TS forms.

Population Screening

Scaling of screening efforts to larger populations serves a number of noteworthy purposes. Foremost, minor-frequency variants present themselves in increasingly larger populations. This point finds significance in the discovery of new variants as part of an otherwise routine process of accumulating data to establish associations and comparative reference levels. Regarding associations, relationships between variants and subgroups in the human population become apparent only when viewed in larger populations. Along similar lines, the presence/absence and modulation of variants linked to subgroups begin to define themselves with some statistical significance in larger populations, forming the basis for comparative frequencies and reference levels.

To illustrate, we screened a mixed population of 162 healthy individuals, 183 individuals with type 2 diabetes mellitus (T2DM) with or without cardiovas-

cular disease (CVD), and 56 individuals with cancer (pancreatic, breast, prostate, and colorectal) as described above. Spectral data were converted to tabular and graphic data to show the presence/absence of each variant, qualified using $S/N > 5$, and their relative abundances, measured by integrating and normalizing ion signals. Another allelic variant and an additional 6 low-frequency posttranslational modifications were evident in the cohort of individuals described above (Table 1). Fig. 2 shows relative abundance data for the 4 GcG variants identified above. When viewed in this manner, associations within the populations become apparent. Foremost, there is a tradeoff between the relative abundance of the unmodified and trisaccharide forms that depends on the presence of the Gc2 allele. Individuals not having the Gc2 allele exhibit a relatively higher abundance of trisaccharide form, heterozygous Gc2 individuals exhibit lower abundances, and in homozygous Gc2 individuals the trisaccharide is not readily evident. A second trend is an increased abundance of the glycated form in the T2DM poor maintenance and T2DM/CVD subgroups. Finally, the monosaccharide form is slightly increased in certain disease subgroups (32).



Data can be further dissected in terms of the frequency at which a variant occurs in relationship to other variables contained in the data. Fig. 3A shows the frequency distribution of 9 variants in their respective cohorts (independent of genotype). Further extraction of these frequency data yields the subgroups shown in

Fig. 3B and C, i.e., the frequency with which the variants occur relative to the major genotypes. Fig. 3B is the data extraction of unmodified GcG, showing a relationship between homozygous Gc1F genotype and the frequency in the T2DM/CVD cohort (arrow). Fig. 3C is the extraction of the second set of columns [GcC(TS)],

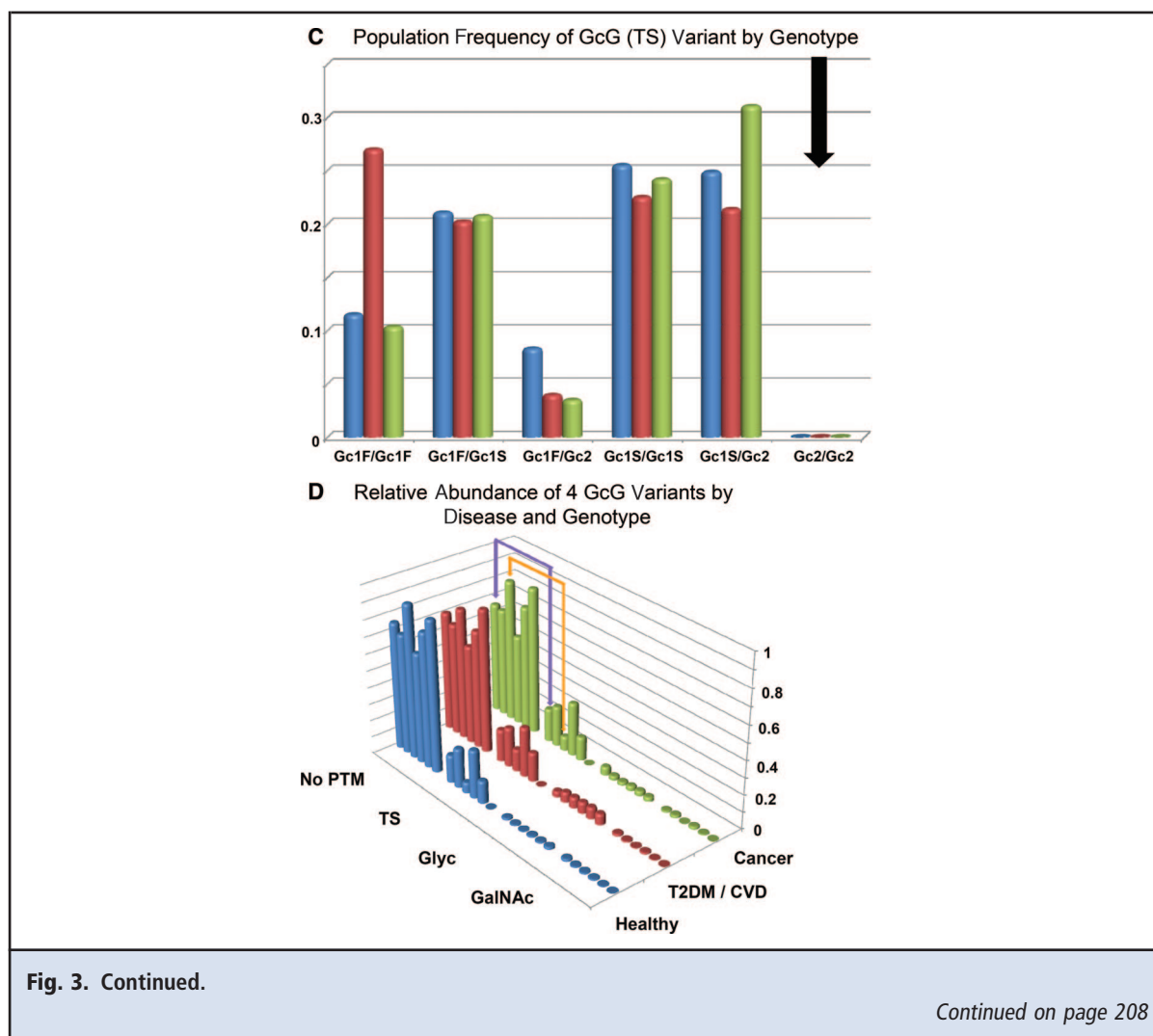


Fig. 3. Continued.

Continued on page 208

which exhibit a lower, but constant, frequency between the cohorts. As indicated by the arrow, the trisaccharide form is not detected in homozygous Gc2 individuals, consistent with the Gc2 genotype lacking the preferred site of *O*-linked glycosylation (Thr⁴²⁰ changed to Lys⁴²⁰) (27, 30).

We generated these frequency views by using the rule that a variant is present if detected at $S/N > 5$. This threshold can be set at other values, or using other metrics, depending on the question posed during extraction. For instance, the unmodified forms were detected in all spectra at $S/N > 300$. This setting is appropriate if the data are queried for genotype alone; i.e., the data shown in Fig. 3B are equivalent to population genotyping data (frequency) and can be merged and/or compared automatically with genomic databases. All other variant data, however, are unique and are able to pro-

duce composite “protein phenotype” views of the cohorts by considering the analog nature of the variant signals. As opposed to frequency counting, which produces a strictly yes or no result, the relative integrals can be viewed in relation to disease, genotype or other variant to add additional depth to the data. For example, in Fig. 3D the cohort data for the 4 variants are subtyped into diploid genotype. Instead of frequency, the y -axis now reports the cohort mean of the normalized integral values. This approach takes into account the frequency of occurrence and relative abundance of the variants as they occur in each cohort, and reports a mean value. The tradeoff between the unmodified and trisaccharide variants is again observed to be linked to Gc2 allele products. Rescaling the GcG (Glyc) section produces Fig. 3E. Relative to healthy individuals, the mean cohort glycation is observed to increase mark-

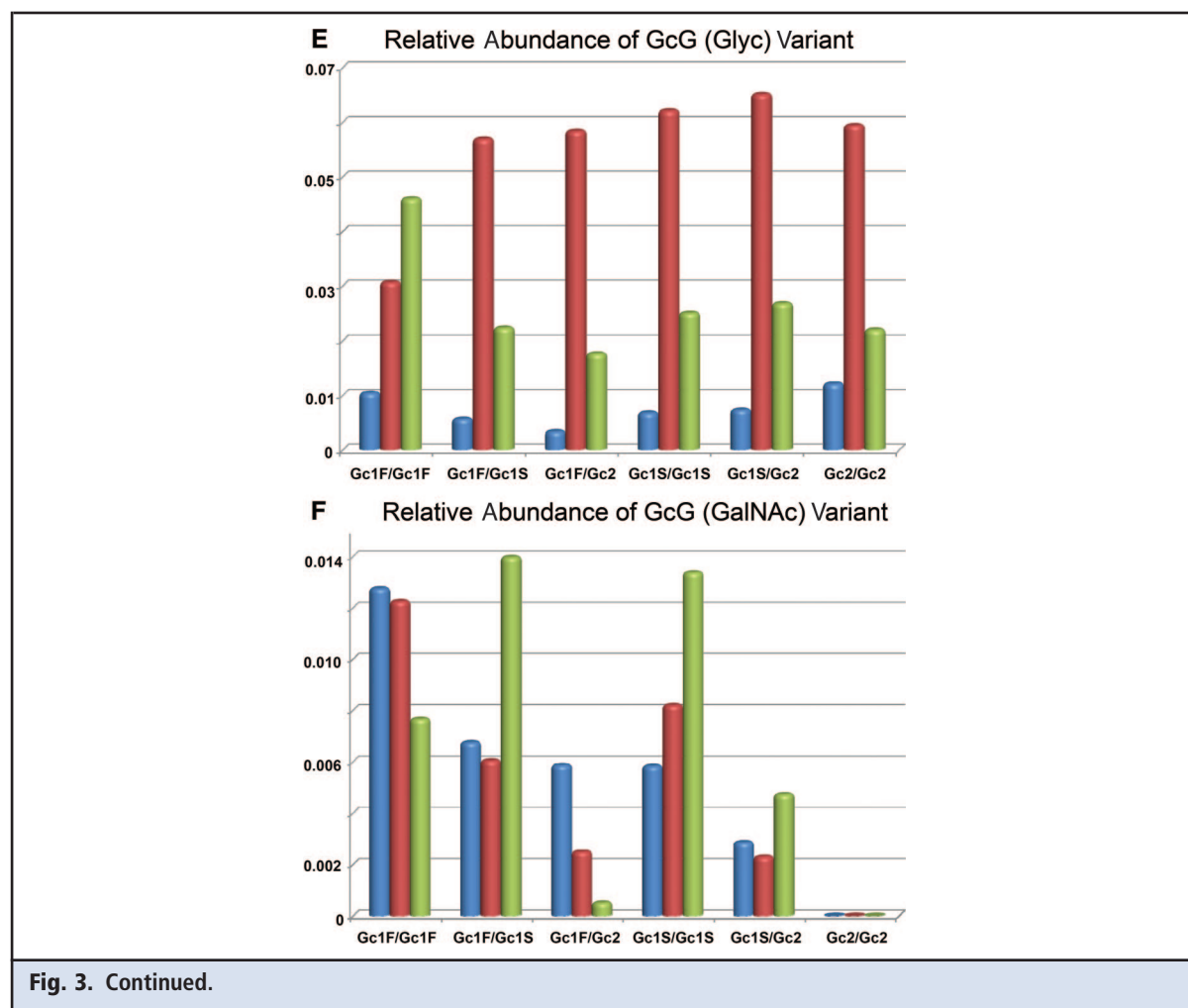


Fig. 3. Continued.

edly across all genotypes in the T2DM poor maintenance and T2DM/CVD cohorts. However, the glycation abundances for the Gc1F/Gc1F subgroup are approximately half those observed for the other genotypes. Considering the higher frequency of Gc1F/Gc1F genotypes observed in the T2DM cohort, this apparently lower relative abundance warrants further investigation to determine possible genotype-dependent reference values in people with diabetes. Additionally, increased glycation is observed in the cancer cohort, suggesting mild hyperglycemia in these individuals. Data for the monosaccharide (GalNAc) variant are shown in Fig. 3F. The relative abundance of the variant is essentially equivalent between the healthy and T2DM cohorts but exhibits a different pattern in the cancer cohort (32). GcG modified by a single GalNAc residue is reportedly a potent macrophage-activating factor (GcMAF) (33), the presence of which is controversial in cancer patients (32, 34). Our data demonstrate that this molecular species is present at very low

concentrations, while indicating a tentative relationship between abundance and Gc1S genotypes.

Applications in Biomarker Development and Clinical Proteomics

The investigation of protein diversity finds importance in biomarker development and clinical proteomics for several reasons. Foremost, a biomarker of disease is a “characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” (definition from the US Food and Drug Administration). Regarding proteins, these characteristics are changes in concentration, amino acid sequence, and/or posttranslational modifications, the relationship of which literally defines microheterogeneity. At its worst, indeterminate microheterogeneity encountered throughout the general population represents biological “noise” that may have

a negative technical impact on the analysis of a biomarker, potentially contributing to the mechanical failure of a conventional assay or producing otherwise misleading results (35). On the other hand, slightly different forms of nominally the same protein are often the upstream cause or downstream effect of alternate *in vivo* activities. Defining and quantifying these events is therefore critical to understanding the pathobiologies underlying diseases. As such, determinate microheterogeneity related to such functionality (or dysfunctionality) behind diseases is an open avenue of investigation to find and develop new biomarkers.

Our exemplar case begins to illustrate these nuisances. Presently, there are no commercial assays able to simultaneously differentiate multiple GcG variants. Existing assays either group all forms into a single value or, if designed for only 1, produce a value for that particular form and miss the others. As the data given above illustrate, intraprotein and disease-specific associations appear when the variants are considered. With knowledge of the molecular identity of the species, trends can be equated to functionality and/or disease pathobiologies, such as increased macrophage-activating factor in cancer (32), genomic knockout of a glycosylation site, and increased glycation in diabetes. Particularly intriguing during biomarker development and clinical proteomics is that molecular variants relevant to 2 different diseases are observed during the analysis of products from a single gene. This phenomenon is not without precedence in the diagnostic world, as exemplified by the routine screening of 2 different molecular forms of hemoglobin (b-chain) in detecting sickle cell anemia (HbS) or monitoring diabetes (HbA1C). The full extent that such phenomena occur on other proteins, for example different protein phenotypes carried on a common protein scaffold, remains to be fully investigated, but the prospects are exciting.

Comments and Future Directions

The investigations described and referred to here begin by designing and validating assays centered on the use of “good” antibodies, defined by their ability to capture peptides or proteins (with and without microheterogeneity) with high selectivity, affinity, and avidity. After 15 years of practice, this task is still undertaken empirically—however, not for lack of available antibodies [there are thousands available commercially, industrially, and through original equipment manufacturer (OEM) sources]. Assay design has evolved to the use of polyclonal or monoclonal antibody cocktails recognizing different epitopes in the target, and more recently to the use of alternate affinity reagents such as aptamers. When chosen judiciously, assays can be pro-

duced with limits of detection and absolute quantification at low pg/mL concentrations, while still detecting (and discovering) microheterogeneity in clinically relevant proteins (35).

Regarding mass spectrometry, it is important to understand that these analyses can be performed in numerous ways (for a comprehensive review, see (36)), and here we offer some simple observations for guidance. In general, MALDI-TOF instrumentation is a rapid and sensitive multiplex approach that is able to analyze several targeted proteins and their variants simultaneously. Quantification always requires the use of multiple signals for intersample normalization, using either the ratio of targeted proteins or intentionally introduced internal standards (14). Instruments based on electrospray ionization are ideal for coupling to liquid chromatography, which can enable multiplexing. Generally, they exhibit adequate performance to resolve numerous posttranslational modifications on proteins as large as immunoglobulins (37). Although standards for quantification are strongly suggested, some instrumental configurations can be used without standards to gauge relative responses between samples. In either approach, the semiquantitative dynamic range (i.e., the accurate measurement of the ratio between 2 signals) in full-scan analyses is between 100 and 1000. Notably, ratios measured using 1 approach are not necessarily the same for an alternate approach, and care should be taken to understand and normalize differences between platforms. Last, it is advantageous to use multiple approaches and platforms in combination when producing high-performance mass spectrometry-based assays for clinically significant proteins (38). The simplest test of the validity of an approach is whether it can be converted to standardized assays and operating procedures and applied to large populations to economically generate useful and reproducible data.

Universally, the value of this form of targeted, full-length characterization of proteins in populations should be self-evident, but it is not presently required during proteomics investigations where even the simplest microheterogeneity can be influential on the commonly used surrogate assays. Surprisingly, the approach is often criticized based on the generalization that antibodies will not capture “all” variant forms, and on the presumed “biased” nature of the analysis. These notions lead to speculations regarding variants escaping detection and are erroneously equated with an inability to discover a new biomarker or molecular species. Conveniently overlooked is that the majority of variants we observe are routinely missed, or not even considered, using conventional proteomics and/or immunometric approaches, and we oftentimes discover variants that have yet to be reported in the literature,

such as 4 of the GcG variants listed in Table 1. These comments are not meant to be inflammatory, but instead to establish a pragmatic foothold regarding a recent initiative calling for the formation of a human proteome detection and quantification project (hPDQ), which aims to produce antibodies targeting peptide/protein epitopes resulting from all genes in the human genome (39). These prophetic antibodies form the basis for standardized biomarker development using immunoaffinity capture in combination with mass spectrometric analysis, much as described herein using available antibodies. Ironically, the technical workflows proposed for the hPDQ virtually guarantee that microheterogeneity will not be detected during routine analyses. At a minimum, the hPDQ will benefit by using data from studies such as described here to assist in designing assays. At the same time, select sets of hPDQ antibodies can be tailored for full-length analysis, or to specifically target determinant microheterogeneity related to disease, and used in population-based screening and biomarker development.

Thus, as part of ongoing biomarker discovery and clinical proteomics initiatives, it may be worthwhile to consider increasing efforts toward the full-length characterization of proteins in large and diverse populations—efforts that go toward addressing biological variability in humans. Other than representing possible molecular confounders that lie in wait as the current technical variables of proteomics are minimized, such diversity is an open avenue of investigation to increase our behavioral knowledge of human proteins. This example, investigating products from only a single gene, begins to illustrate the breadth of microheterogeneity possible in humans and gives a few rudimentary means of understanding its ramifications and uses. With the exception of the genotype frequency, the population data are unique and illustrate a need to

more fully understand the exact idiosyncrasies of individual proteins in relation to health and disease. Expanding this concept leads to the proposition of creating a protein diversity database, in which protein diversity is indexed relative to disease, stage, treatments, age, sex, race, geographic region, and other epidemiological metrics, and that contains baseline information on a sizable healthy population. Other than through isolated efforts (40), large-scale proteomics investigations have yet to produce such a database that tracks protein behavior in populations. It is easy, however, to anticipate its value in defining technically confounding variants, assisting in the design of standardized assays, aiding in elucidating protein function, building models of disease-centric networks and pathways, and translating new disease markers and technologies closer to clinical and diagnostic applications.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors' Disclosures of Potential Conflicts of Interest: No authors declared any potential conflicts of interest.

Acknowledgments: Funding was provided by the Arizona State University's Technology and Research Initiative Fund, the Agilent Technologies Foundation (Research Project Gift 08US-422UR), and the NIH (grant numbers R21RR024440 and R24DK083948). There were no conflicts of interest for this paper.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Disclaimers: The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or other funding agencies.

References

- Nelson RW, Krone JR, Bieber AL, Williams P. Mass spectrometric immunoassay. *Anal Chem* 1995;67: 1153–8.
- Kiernan UA, Addobbati R, Nedelkov D, Nelson RW. Quantitative multiplexed C-reactive protein mass spectrometric immunoassay. *J Proteome Res* 2006;5:1682–7.
- Kiernan UA, Nedelkov D, Nelson RW. Multiplexed mass spectrometric immunoassay in biomarker research: a novel approach to the determination of a myocardial infarct. *J Proteome Res* 2006;5: 2928–34.
- Kiernan UA, Tubbs KA, Gruber K, Nedelkov D, Niederkofer EE, Williams P, Nelson RW. High-throughput protein characterization using mass spectrometric immunoassay. *Anal Biochem* 2002; 301:49–56.
- Niederkofer EE, Tubbs KA, Gruber K, Nedelkov D, Kiernan UA, Williams P, Nelson RW. Determination of beta-2 microglobulin levels in plasma using a high-throughput mass spectrometric immunoassay system. *Anal Chem* 2001;73: 3294–9.
- Niederkofer EE, Tubbs KA, Kiernan UA, Nedelkov D, Nelson RW. Novel mass spectrometric immunoassays for the rapid structural characterization of plasma apolipoproteins. *J Lipid Res* 2003;44:630–9.
- Tubbs KA, Nedelkov D, Nelson RW. Detection and quantification of beta-2-microglobulin using mass spectrometric immunoassay. *Anal Biochem* 2001; 289:26–35.
- Nedelkov D. Mass spectrometry-based immunoassays for the next phase of clinical applications. *Exp Rev Proteom* 2006;3:631–40.
- Nedelkov D, Tubbs KA, Niederkofer EE, Kiernan UA, Nelson RW. High-throughput comprehensive analysis of human plasma proteins: a step toward population proteomics. *Anal Chem* 2004;76: 1733–7.
- Nelson RW, Nedelkov D, inventors; Intrinsic Bio-probes, Inc., assignee. Bioactive chip mass spectrometry. US patent 6,887,713. 2005 May 3.
- Nelson RW, Nedelkov D, Tubbs KA, Kiernan UA. Quantitative mass spectrometric immunoassay of insulin like growth factor 1. *J Proteome Res* 2004; 3:851–5.
- Kiernan UA, Tubbs KA, Nedelkov D, Niederkofer EE, McConnell E, Nelson RW. Comparative urine protein phenotyping using mass spectrometric immunoassay. *J Proteome Res* 2003;2:191–7.
- Kiernan UA, Tubbs KA, Nedelkov D, Niederkofer EE, Nelson RW. Detection of novel truncated forms of human serum amyloid A protein in human plasma. *FEBS Lett* 2003;537:166–70.

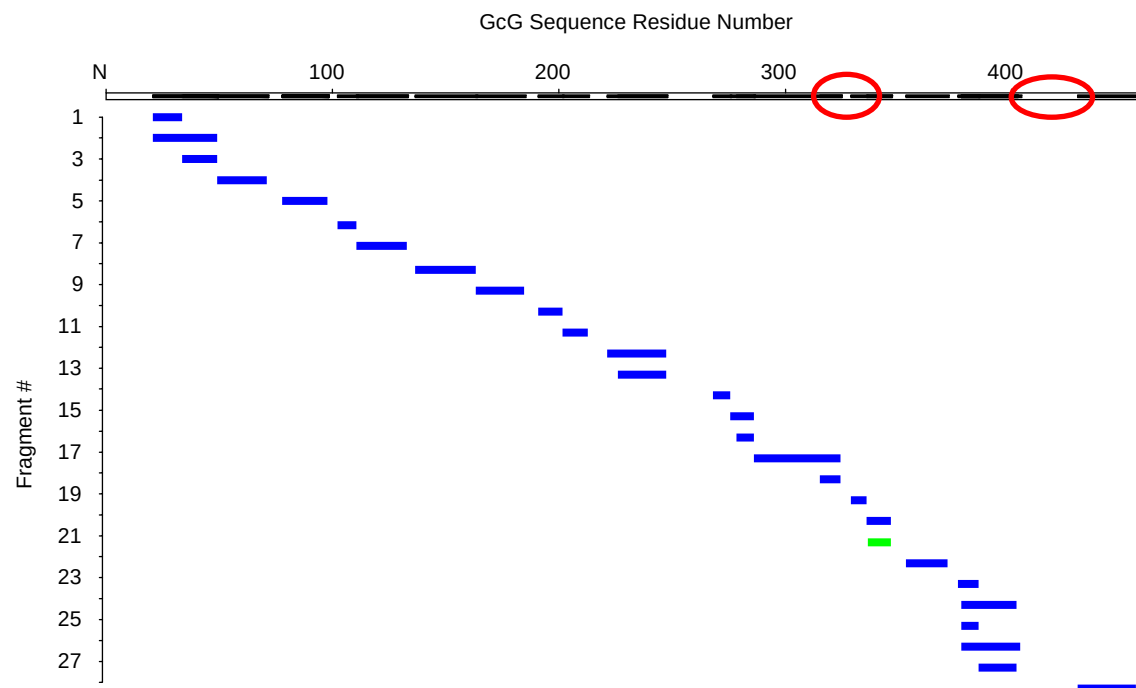
14. Nelson RW, McLean MA, Hutchens TW. Quantitative determination of proteins in solution by matrix-assisted laser desorption time-of-flight mass spectrometry. *Anal Chem* 1994;66:1408–15.
15. Nedelkov D, Kiernan UA, Niederkofler EE, Tubbs KA, Nelson RW. Investigating diversity in human plasma proteins. *Proc Natl Acad Sci U S A* 2005;102:10852–7.
16. Nedelkov D, Phillips DA, Tubbs KA, Nelson RW. Investigation of human protein variants and their frequency in the general population. *Mol Cell Proteomics* 2007;6:1183–7.
17. Aebersold R. A stress test for mass spectrometry-based proteomics. *Nature Methods* 2009;6:411–2.
18. Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, et al. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nature Methods* 2009;6:423–30.
19. Prakash C, Shaffer CL, Nedderman A. Analytical strategies for identifying drug metabolites. *Mass Spectrom Rev* 2007;26:340–69.
20. Maurer HH. Multi-analyte procedures for screening for and quantification of drugs in blood, plasma, or serum by liquid chromatography-single stage or tandem mass spectrometry (LC-MS or LC-MS/MS) relevant to clinical and forensic toxicology. *Clin Biochem* 2005;38:310–8.
21. Maurer HH. Current role of liquid chromatography-mass spectrometry in clinical and forensic toxicology. *Anal Bioanal Chem* 2007;388:1315–25.
22. Ma S, Chowdhury SK, Alton KB. Application of mass spectrometry for metabolite identification. *Curr Drug Metab* 2006;7:503–23.
23. Addona TA, Abbatiello SE, Schilling B, Skates SJ, Mani DR, Bunk DM, et al. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat Biotechnol* 2009;27:633–41.
24. Kuzyk MA, Smith D, Yang J, et al. Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. *Mol Cell Proteomics* 2009;8:1860–77.
25. Duncan MW, Yergey AL, Patterson SD. Quantifying proteins by mass spectrometry: the selectivity of SRM is only part of the problem. *Proteomics* 2009;9:1124–7.
26. Hancock WS, Wu SL. Human proteinpedia. http://www.humanproteinpedia.org/Experimental_details?exp_id=TE-99440,TE-99889. (Accessed July 2009).
27. Speeckaert M, Huang G, Delanghe JR, Taes YE. Biological and clinical aspects of the vitamin D binding protein (Gc-globulin) and its polymorphism. *Clin Chim Acta* 2006;372:33–42.
28. Fang Y, van Meurs JB, Arp P, van Leeuwen JP, Hofman A, Pols HA, Uitterlinden AG. Vitamin D binding protein genotype and osteoporosis. *Calcif Tissue Int* 2009;85:85–93.
29. Harrison HH, Ober C, Miller KL, Elias S. High-resolution two-dimensional electrophoretic survey of serum protein genetic types in Schiedeleut Hutterites. *Am J Human Biol* 1991;3:639–46.
30. Borges CR, Jarvis JW, Oran PE, Nelson RW. Population studies of vitamin D binding protein microheterogeneity by mass spectrometry lead to characterization of its genotype-dependent O-glycosylation patterns. *J Proteome Res* 2008;7:4143–53.
31. Borges CR, Jarvis JW, Oran PE, Rogers SP, Nelson RW. Population studies of intact vitamin D binding protein by affinity capture ESI-TOF-MS. *J Biomol Tech* 2008;19:167–76.
32. Rehder DS, Nelson RW, Borges CR. Glycosylation status of vitamin D binding protein in cancer patients. *Protein Sci* 2009;18:2036–42.
33. Yamamoto N. Structural definition of a potent macrophage activating factor derived from vitamin D3-binding protein with adjuvant activity for antibody production. *Mol Immunol* 1996;33:1157–64.
34. Yamamoto N, Naraparaju VR, Asbell SO. Deglycosylation of serum vitamin D3-binding protein leads to immunosuppression in cancer patients. *Cancer Res* 1996;56:2827–31.
35. Niederkofler EE, Kiernan UA, O'Rear J, Menon S, Saghir S, Protter AA, et al. Detection of endogenous B-type natriuretic peptide at very low concentrations in patients with heart failure. *Circ Heart Fail* 2008;1:258–64.
36. Watson JT, Sparkman OD. Introduction to mass spectrometry: instrumentation, applications, and strategies for data interpretation. 4th ed. West Sussex (UK): Wiley & Sons; 2007.
37. Gadgil HS, Bondarenko PV, Pipes G, Rehder D, McAuley A, Perico N, et al. The LC/MS analysis of glycation of IgG molecules in sucrose containing formulations. *J Pharm Sci* 2007;96:2607–21.
38. Lopez MF, Rezai T, Sarracino DA, Prakash A, Krastins B, Athanas M, et al. Selected reaction monitoring-mass spectrometric immunoassay responsive to parathyroid hormone and related variants. *Clin Chem* 2010;56:281–90.
39. Anderson NL, Anderson NG, Pearson TW, Borchers CH, Paulovich AG, Patterson SD, et al. A human proteome detection and quantitation project. *Mol Cell Proteomics* 2009;8:883–6.
40. Costello CE. TTR database. <http://www.bumc.bu.edu/msr/ttr-database>. (Accessed July 2009).

Supplemental Data Figure Captions

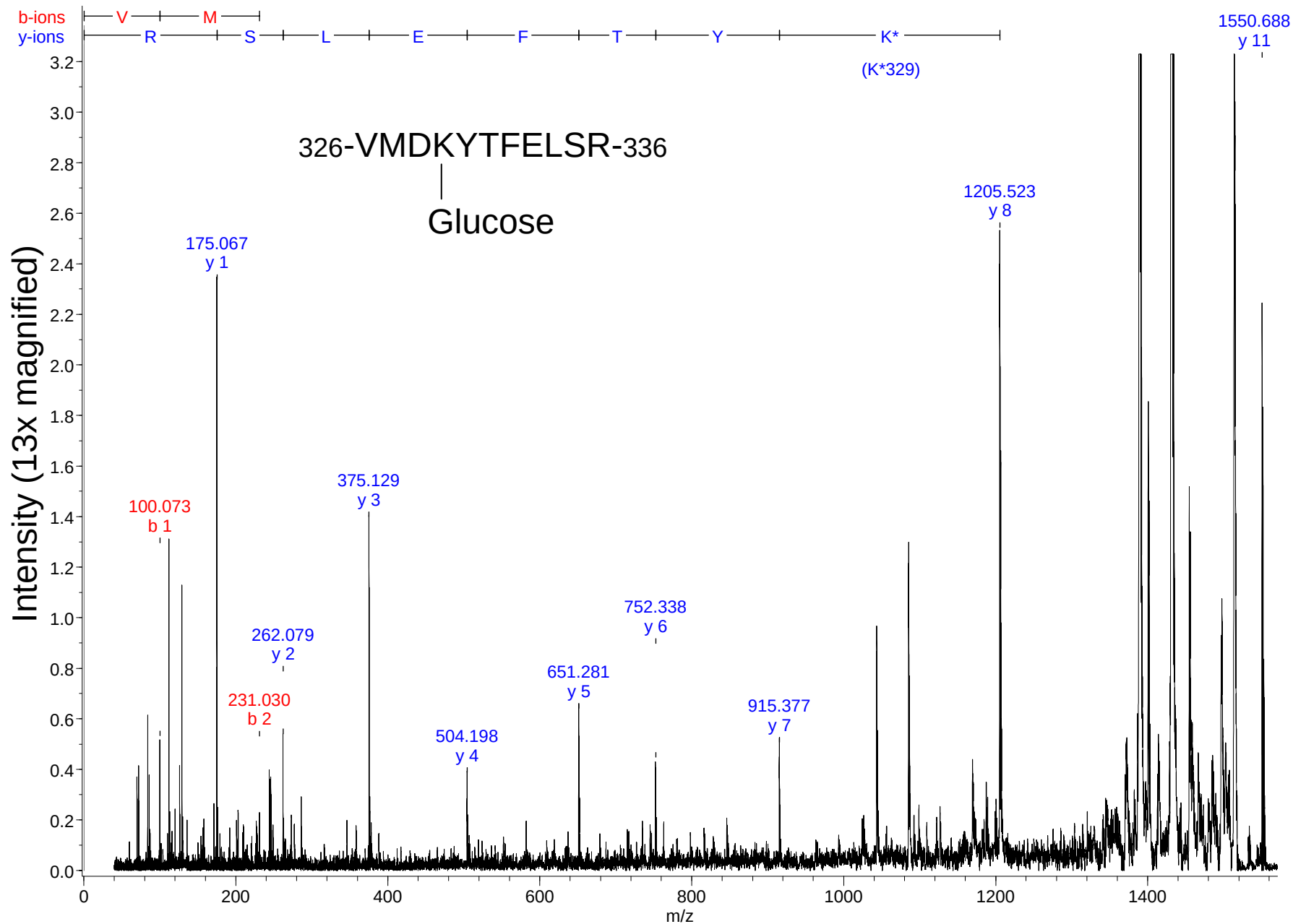
Supplemental Data Fig. 1. Public domain surrogate proteomics data from human blood plasma mapped to GcG sequence. Twenty-eight tryptic fragments are found to correspond with GcG sequence [(28); blue). Tryptic fragment 338-347 is used for quantification [(26); green). Notably, 30 – to – 98% of the sequence is not sampled, respectively, using the surrogate approaches. Gaps in regions found to be heterogeneous (residues 329, and residues 416-420) are denoted with circles.

Supplemental Data Fig. 2. Tandem mass spectra verifying sites and identities of two commonly observed posttranslational modifications of GcG. Data were acquired after enzymatic digestion of GcG found to have prominent signals of the variant forms. **(A)** MALDI-LIFT-TOF/TOFMS spectrum of the precursor ion m/z 1550.74, representing the glycosylated form of a tryptic GcG peptide spanning residues Val³²⁶ – Arg³³⁶. The spacing of 290.15 Da between the y7 and y8 ions indicates that the glucose residue is attached to Lys³²⁹ (128.09 Da for a Lys residue plus 162.05 Da for a glucose residue give a calculated mass shift of 290.14 Da). All masses in this spectrum were within 0.2 Da of the calculated masses. **(B)** Spectrum of the disaccharide (Gal₁GalNAc₁)-modified form of an Arg-C generated GcG peptide spanning residues Leu⁴¹⁰ – Arg⁴²⁹ obtained from a mixture of Gc1F and Gc1S allele products. The precursor ion selection window was widened to allow for peptides from both allele products to undergo analysis by MS/MS. A y11 ion at m/z 1268.18 (calc. 1268.17) suggests that a portion of the Gal₁GalNAc₁ disaccharide is attached at Thr⁴¹⁸. An additional signal at m/z 1633.04 (calc. 1633.09),

corresponds to an alternate y11 ion indicating an additional attachment site of the disaccharide at Thr⁴²⁰. See references ([30-32](#)) for additional details regarding site-specific glycosylation of GcG. Asterisks indicate residues that are supported by the data as sites of modification.



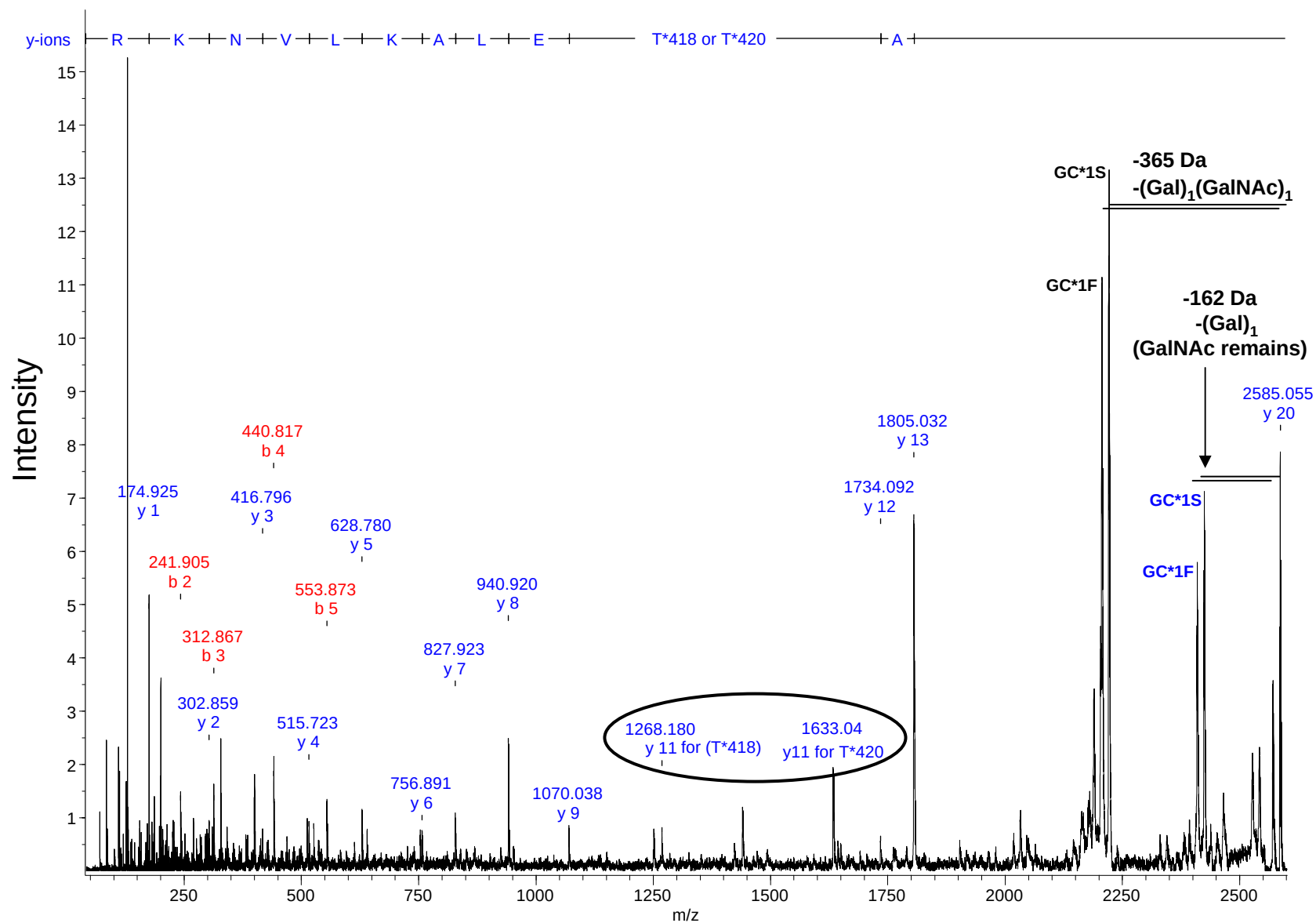
Supplemental Data Figure 1



Supplemental Data Figure 2A

GC*1S 410-LKAKLPEATPTTELAKLVNKR-429

GC*1F $\updownarrow$ D $\updownarrow$ OR $\updownarrow$ GalNAc₁Gal₁ (attached to T418 in some cases but T420 in others—see y11 ions)



Supplemental Data Figure 2B