Circulation

EDITORIAL

In Aptamers They Trust

Caveats of the SOMAscan Biomarker Discovery Platform From SomaLogic

Article, see p 2469

Abhishek Joshi, BA (Hons), BMBCh
Manuel Mayr, MD, PhD

Although it has become affordable to sequence the entire human genome, measuring all proteins in a sample as complex as human plasma remains currently out of reach. The plasma proteome has an enormous dynamic range and variability, including splice variants, cleavage products, and posttranslational modifications. Antibody-based techniques have predominated, but other affinity-based techniques such as the SOMAscan aptamer assays from SomaLogic also promise the multiplexed measurement of proteins in a scalable manner.

Aptamers are short oligonucleotides that have binding affinity to a single protein. They have been developed by introducing a pool of random sequence oligomers to a target protein, applying affinity selection to separate target-oligomer pairs, and then amplifying the surviving sequences. Iterations of this cycle, called SELEX (Systematic Evolution of Ligands by Exponential Enrichment), increase the specificity and avidity of the surviving oligomers to a target protein, eventually identifying a single oligomer, an aptamer, with high affinity for a protein epitope. Additions of side chains improve the stability of aptamers in biological matrices such as plasma. These modifications also change the binding characteristics of the aptamer, giving the same sequence many different properties. The resulting SOMAmers (Slow Off-Rate Modified Aptamers) can be bound to fluorophores and multiplexed, potentially allowing simultaneous quantification of hundreds to thousands of proteins. The most recent iteration of the platform offers ≈5000 aptamers. The SOMAscan v1.3 platform consists of 1129 aptamers.[1]

In the current edition of *Circulation*, Mosley and colleagues[2] use the SOMAscan v1.3 platform in the Framingham Offspring Cohort and report plasma levels of 268 of the measured 1129 proteins to be regulated by single nucleotide polymorphisms (SNPs). Robustness of the findings was demonstrated in the Cooperative Health Research in the Region of Augsburg (KORA) F4 data set, using the same platform. When tested in a third cohort (MDCS [Malmö Diet and Cancer Study]) the strongest gene-protein instruments predicted 60% of a protein's variance, although the median prediction was 8% of the total variance. These gene-protein instruments were used to impute the protein levels of 41 288 previously genotyped individuals in the eMERGE cohort (Electronic Medical Records and Genomics Network), generating the virtual proteome, and their associations with 1128 clinical phecodes derived from electronic health records (EHRs) were assessed. Fifty-five virtual proteins were found to associate with 89 clinical diagnoses in a complex manner; some proteins were associated with more than 1 diagnosis (for example, factor XI was associated with 4 different diagnoses related to venous thrombosis). Some conditions had many associated proteins (for example, thrombosis was predicted by 16 different protein levels). Finally, proteins that were virtually associated

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

Key Words: Editorials ■ aptamers, nucleotide ■ biomarkers ■ oligonucleotides ■ polymorphism, single nucleotide ■ proteomics

https://www.ahajournals.org/journal/circ

with disease outcomes were validated by independent measurements of either elevated low-density lipoprotein cholesterol or ultrasound-confirmed atherosclerosis. Direct protein measurements were made for the 7 gene-protein instruments that virtually associated with elevated low-density lipoprotein cholesterol in MCDS and were found to explain a median variation of 3% of these proteins' abundance. Four of these proteins were isoforms of apolipoprotein E. The other 3 proteins (catalase, interleukin-27, and granulin) did not survive independent validation. When focusing on atherosclerosis, only 2 of the 6 associated virtual proteins could be validated: CLC1B (C-type lectin domain family 1 member B) and PDGFR-ß (platelet-derived growth factor receptor ß).

A potential strength of this study is the achievement of high-throughput, scaled data collection and analysis. The authors achieve this in 3 scaling steps. The first is to use the SOMAscan platform to make the initial protein measurements. The second is to impute protein results in a larger cohort using prediction instruments to generate the virtual proteome. This leverages the huge amount of readily available genomic data and sidesteps the requirement to measure and analyze individual samples, with its attendant complexities of adequate sample quality, storage, preparation, and analysis. Finally, the use of EHRs to generate phecodes allows comparisons across many disease processes and clinical diagnoses, and effectively multiplexes the diseases that can be analyzed.

The first scaling step relies on the specificity, reproducibility, and quantitative accuracy of the SOMAscan aptamer assays. This platform uses a single binder and direct readout that can jeopardize specificity (Table). In an early assessment using chicken plasma as a negative control, 27.6% of SOMAmers (312/1129) generated signals 10-fold those of human plasma.[4] A few SOMAmer-protein–bound complexes have been characterized, but most target epitopes remain mysterious. A single aptamer may have more than 1 target, especially in a complex matrix such as plasma. It is crucial to note that alterations to the protein structure attributable to oligomerization, degradation, posttranslational modifications, or genetic polymorphisms may significantly but unpredictably alter binding affinity and quantification. Another challenge is the potential for batch or plate effects after normalization strategies are applied. Previous investigators note the use of more than one calibrator batch across large cohorts, and normalization strategies that lead to potential intraplate effects.[5] The SOMAscan assay is designed as a discovery platform and measures relative protein concentrations using only external controls. Without internal controls and standard curves, it remains unclear which measurements are within the linear dynamic range.

The second scaling method in this study is the prediction of the virtual proteome in the eMERGE population, based on the association of SNPs with SOMAmer-derived variation in the discovery cohorts. This imputation strategy allows the approximation of the plasma proteome at a scale not available to empirical investigators. A similar approach has been described recently in the INTERVAL study (To Determine Whether the Interval Between Blood Donations in England Can be Safely and Acceptably Decreased); an expanded SOMAscan panel (3622 protein measurements) identified 1927 regulatory SNPs for protein levels.[3] This study demonstrated that 14% of aptamers showed nonspecific binding; 7% bound more than 1 protein, and the remaining 7% bound a protein isoform. Amino acid substitutions conferred by SNPs significantly alter the affinity of aptamers for their target proteins in 32% of the panel, meaning that variations in the predicted abundance assigned to these SNPs are possibly artefactual. When validation using an antibody-based platform was performed in the study by Sun and col-

Table. Description of Possible Sources of Errors in Quantification of Protein Abundance in the SOMAscan Platform

Cause of Interference	Mechanism	Effect
Cross-reactivity	Aptamers can bind with high affinity either to nontarget human proteins, or to different isoforms of the same human protein	7% bind to a different protein 7% bind to an isoform[3]
Nonspecificity	Aptamers are not specific to human proteins but bind to proteins from other species	27.6% of aptamers detect proteins in chicken plasma with 10-fold greater signal intensity than in human plasma[4]
Sequence variation	Single amino acid changes in the aptamer binding site can alter the affinity of binding, and thus spuriously alter the measured abundance	32% of aptamers display altered binding affinity attributable to SNPs[3]
Batch effects	Because aptamers do not provide absolute quantification, calibrators are required to ensure a constant measurement across each plate, and normalization strategies are applied	Intraplate coefficient of variability can be as high as 3.6%, similar to the size of some reported genetic effects[5]
Protein complexes	Plasma proteins can form complexes that may alter their tertiary structure and aptamer affinity	To be determined
DNA/RNA-binding proteins	DNA/RNA-binding proteins[6] and other carrier proteins, such as albumin, may differentially sponge aptamers in plasma and compromise accurate quantitation	To be determined

SNP indicates single nucleotide polymorphism.

leagues,[3] 35% of SNP-protein predictive instruments were not replicated. The authors of the current study acknowledge this limitation; a high false-negative rate is the price of statistical prudence, but false positives in studies such as these can undermine or misdirect future research. Beyond the current study, Sun and colleagues' findings raise concerns not only about the validity of correlating SNPs with this method of affinity-based protein measurement, but also for the wider use of a technique that could give spurious results where minor allele frequencies approach 5%.

The final scaling strategy, the use of EHR, effectively allows the investigators to interrogate over a thousand clinical diagnoses, increasing the potential rate of biomarker discovery. Proteomic investigations previously have only been able to interrogate the prespecified target disease. The use of EHR in epidemiological studies is in its infancy, but has already been used, for example, to predict bleeding outcomes during antiplatelet therapy.[7] However, EHR diagnosis codes are not collected with the aim of providing robust scientific data. In the United States, where the eMERGE cohort is recruited, EHR codes are used for billing and costing services. In the United Kingdom, the same codes are used by primary care physicians to document medical diagnoses, in part, to meet incentivized treatment areas. In both cases, reporting biases would be expected, but have not yet been adequately studied or quantified to allow claims about their accuracy.[8,9] There is currently no consensus on the accuracy of EHRs, but biases will likely include underdiagnosis of rare diseases and overreporting of diseases in incentivized areas of medical practice. Until the landscape is better mapped, studies using phecodes as surrogates for outcomes will remain speculative.

The integrated use of genomic, proteomic, and phenomic data in combination with empirical and statistical methodology could change the way biomarkers are discovered. A similar aptamer-based approach in a prospective analysis of acute cardiovascular events in a cohort with stable coronary artery disease identified a group of 9 proteins, including troponin I and matrix metalloproteinase-12, that modestly increased the predictive power of a clinical risk score.[10] These are already well-described biomarkers. The current study faces the same challenge; some of the protein biomarkers that withstand adjustment for the multiple statistical analyses are already well known, whereas others are missing. Apolipoprotein E levels have been previously identified as a biomarker for cardiovascular risk by direct measurement with mass spectrometry, the gold standard for specific protein identification, which avoids reliance on affinity-based techniques.[11–13] However, other well-known SNP-regulated plasma proteins were not detected. For example, apolipoprotein B is strongly related to cardiovascular outcomes but apparently missing from the analysis, despite being included in the SOMAscan panel. These missing findings do not invalidate the positive results presented in this study, but highlight how many other disease associations the genetically predicted protein levels based on the aptamer platform might miss. Another surprising observation in this study is that almost a quarter of associations between genetic predictors and clinical phenotypes are attributed to thrombosis. Alternative explanations for this overrepresentation of thrombosis might include that a prothrombotic state or preanalytical variations in plasma preparations impact aptamer binding. Finally, the high number of associations with the ABO locus highlight the potential for single SNPs to confer pleiotropic effects, rendering the individual proteins they control less meaningful.[14]

In summary, commercial solutions offer a rapid and convenient way of outsourcing protein measurements, and imputation removes the inconvenience of empirical measurement. Proof of principle, however, still requires independent validation of protein levels through orthogonal validation techniques and linkage to *cis*-acting Mendelian randomization that do not have epitope effects. In particular, changes in electric charge caused by amino acid substitution may alter the binding properties of the negatively charged aptamers. Thus, the SOMAscan may not be as unbiased as it first may seem. The imputation approach may work well for certain plasma proteins that have a high degree of heritability and can be demonstrated to avoid the potential confounding we outline. These protein quantitative trait loci, made publicly available by the authors, could provide the map for further investigations in highly genotyped populations. Overall, the increasing abstraction from empirical science leaves us with a sense of virtual reality; to quote from *Tractatus Logico-Philosophicus* by the Austrian philosopher Ludwig Wittgenstein, "Whereof one cannot speak, thereof one must be silent."[15]

ARTICLE INFORMATION

Correspondence

Manuel Mayr, MD, PhD, King's British Heart Foundation Centre, King's College London, 125 Coldharbour Ln, London SE5 9NU, United Kingdom. Email manuel.mayr@kcl.ac.uk

Affiliations

King's College London British Heart Foundation Centre, School of Cardiovascular Medicine and Sciences, United Kingdom (A.J., M.M.). Bart's Heart Centre, St. Bartholomew's Hospital, London, United Kingdom (A.J.).

Sources of Funding

Dr Joshi is a British Heart Foundation Clinical Research Training Fellow (FS/16/32/32184). Dr Mayr is a British Heart Foundation Chair Holder (CH/16/3/32406) with British Heart Foundation program grant support (RG/16/14/32397).

Disclosures

Dr Mayr is named inventor on patents related to biomarkers in cardiometabolic disease.

REFERENCES

1. Rohloff JC, Gelinas AD, Jarvis TC, Ochsner UA, Schneider DJ, Gold L, Janjic N. Nucleic acid ligands with protein-like side chains: modified aptamers and their use as diagnostic and therapeutic agents. *Mol Ther Nucleic Acids*. 2014;3:e201. doi: 10.1038/mtna.2014.49
2. Mosley JD, Benson MD, Smith JG, Melander O, Ngo D, Shaffer CM, Ferguson JF, Herzig MS, McCarty CA, Chute CG, Jarvik GP, Gordon AS, Palmer MR, Crosslin DR, Larson EB, Carrell DS, Kullo IJ, Pacheco JA, Peissig PL, Brilliant MH, Kitchner TE, Linneman JG, Namjou B, Williams MS, Ritchie MD, Borthwick KM, Kiryluk K, Mentch FD, Sleiman PM, Karlson EW, Verma SS, Zhu Y, Vasan RS, Yang Q, Denny JC, Roden DM, Gerszten RE, Wang TJ. Probing the virtual proteome to identify novel disease biomarkers. *Circulation*. 2018;138:2469–2481. doi: 10.1161/CIRCULATIONAHA.118.036063
3. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, Oliver-Williams C, Kamat MA, Prins BP, Wilcox SK, Zimmerman ES, Chi A, Bansal N, Spain SL, Wood AM, Morrell NW, Bradley JR, Janjic N, Roberts DJ, Ouwehand WH, Todd JA, Soranzo N, Suhre K, Paul DS, Fox CS, Plenge RM, Danesh J, Runz H, Butterworth AS. Genomic atlas of the human plasma proteome. *Nature*. 2018;558:73–79. doi: 10.1038/s41586-018-0175-2
4. Christiansson L, Mustjoki S, Simonsson B, Olsson-Strömberg U, Loskog ASI, Mangsbo SM. The use of multiplex platforms for absolute and relative protein quantification of clinical material. *EuPA Open Proteomics*. 2014;3:37–47. doi:10.1016/J.EUPROT.2014.02.002
5. Candia J, Cheung F, Kotliarov Y, Fantoni G, Sellers B, Griesman T, Huang J, Stuccio S, Zingone A, Ryan BM, Tsang JS, Biancotto A. Assessment of variability in the SOMAscan assay. *Sci Rep*. 2017;7:14248. doi: 10.1038/s41598-017-14755-5
6. Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, Mitchell PS, Bennett CF, Pogosova-Agadjanyan EL, Stirewalt DL, Tait JF, Tewari M. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc Natl Acad Sci USA*. 2011;108:5003–5008. doi: 10.1073/pnas.1019055108
7. Pasea L, Chung SC, Pujades-Rodriguez M, Moayyeri A, Denaxas S, Fox KAA, Wallentin L, Pocock SJ, Timmis A, Banerjee A, Patel R, Hemingway H. Personalising the decision for prolonged dual antiplatelet therapy: development, validation and potential impact of prognostic models for cardiovascular events and bleeding in myocardial infarction survivors. *Eur Heart J*. 2017;38:1048–1055. doi: 10.1093/eurheartj/ehw683
8. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, Shah AD, Timmis AD, Schilling RJ, Hemingway H. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. *PLoS One*. 2014;9:e110900. doi: 10.1371/journal.pone.0110900
9. Wei WQ, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, Cox NJ, Roden DM, Denny JC. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One*. 2017;12:e0175508. doi: 10.1371/journal.pone.0175508
10. Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, Sterling DG, Williams SA. Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA*. 2016;315:2532–2541. doi: 10.1001/jama.2016.5951
11. Pechlaner R, Tsimikas S, Yin X, Willeit P, Baig F, Santer P, Oberhollenzer F, Egger G, Witztum JL, Alexander VJ, Willeit J, Kiechl S, Mayr M. Very-low-density lipoprotein-associated apolipoproteins predict cardiovascular events and are lowered by inhibition of APOC-III. *J Am Coll Cardiol*. 2017;69:789–800. doi: 10.1016/j.jacc.2016.11.065
12. Mayr M, Gerszten R, Kiechl S. Cardiovascular risk beyond low-density lipoprotein cholesterol. *J Am Coll Cardiol*. 2018;71:633–635. doi: 10.1016/j.jacc.2017.12.040
13. Yin X, Baig F, Haudebourg E, Blankley RT, Gandhi T, Müller S, Reiter L, Hinterwirth H, Pechlaner R, Tsimikas S, Santer P, Willeit J, Kiechl S, Witztum JL, Sullivan A, Mayr M. Plasma proteomics for epidemiology: increasing throughput with standard-flow rates. *Circ Cardiovasc Genet*. 2017;10:e001808. doi:10.1161/CIRCGENETICS.117.001808
14. Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, Hoover H, Gudmundsdottir V, Horman SR, Aspelund T, Shu L, Trifonov V, Sigurdsson S, Manolescu A, Zhu J, Olafsson Ö, Jakobsdottir J, Lesley SA, To J, Zhang J, Harris TB, Launer LJ, Zhang B, Eiriksdottir G, Yang X, Orth AP, Jennings LL, Gudnason V. Co-regulatory networks of human serum proteins link genetics to disease. *Science*. 2018;361:769–773. doi: 10.1126/science.aaq1327
15. Wittgenstein L. *Tractatus Logico-Philosophicus*. New York: Routledge Classics; 2001.