Published in final edited form as:
Circulation. 2018 November 27; 138(22): 2482–2485. doi:10.1161/CIRCULATIONAHA.118.036823.

In Aptamers They Trust: The Caveats of the SOMAscan Biomarker Discovery Platform from SomaLogic

Abhishek Joshi, BA (Hons), BMBCh[1,2] and **Manuel Mayr, MD, PhD**[1]

[1]King's College London British Heart Foundation Centre, School of Cardiovascular Medicine and Sciences, London, United Kingdom

[2]Bart's Heart Centre, St. Bartholomew's Hospital, London, United Kingdom

Whilst 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 post-translational 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, which 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 Slow Off-rate Modified Aptamers (SOMAmers) 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 around 5000 aptamers. The SOMAscan v1.3 platform consists of 1129 aptamers[1].

In the current edition of this Journal, Moseley and colleagues 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)[2]. Robustness of the findings was demonstrated in the KORA F4 dataset, using the same platform. When tested in a third cohort (Malmö Diet and Cancer Study or MDCS) the strongest gene-protein instruments predicted 60% of a protein's variance, although the median prediction was only 8% of the total variance. These gene-protein instruments were

Correspondence: Manuel Mayr, MD, PhD. King's British Heart Foundation Centre, King's College London, 125 Coldharbour Lane, London SE5 9NU, UK. manuel.mayr@kcl.ac.uk, phone: +44 20 7848 5446, fax: +44 20 7848 5298, twitter: @vascular_prot.

Disclosures:
M.M. is named inventor on patents related to biomarkers in cardiometabolic disease.

Europe PMC Funders Author Manuscripts

used to impute the protein levels of 41,288 previously genotyped individuals in the eMERGE cohort, generating the “virtual proteome”, and their associations with 1,128 clinical “Phecodes” derived from Electronic Health Records (EHRs) were assessed. 55 virtual proteins were found to associate with 89 clinical diagnoses in a complex manner; some proteins were associated with more than one 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 which were virtually associated with disease outcomes were validated by independent measurements of either elevated LDL-C or ultrasound-confirmed atherosclerosis. Direct protein measurements were made for the 7 gene-protein instruments which virtually associated with elevated LDL-C in MCDS and were found to explain a median variation of 3% of these proteins’ abundance. Four of these proteins were isoforms of ApoE. The other 3 proteins (catalase, interleukin-27 and granulin) did not survive independent validation. When focussing on atherosclerosis, only 2 of the 6 associated virtual proteins could be validated; C-type lectin domain family 1 member B (CLC1B) and platelet-derived growth factor receptor beta (PDGFR-beta).

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 analyse 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 analysed.

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, which can jeopardise specificity (Table 1). In an early assessment using chicken plasma as a negative control, 27.6% of SOMAmers (312/1129) generated signals tenfold that of in human plasma[4]. A few SOMAmer-protein bound complexes have been characterised, but most target epitopes remain mysterious. A single aptamer may have more than one target, especially in a complex matrix such as plasma. Crucially, alterations to the protein structure due to oligomerisation, 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 normalisation strategies are applied. Previous investigators note the use of more than one calibrator batch across large cohorts, and normalisation strategies that lead to potential intra-plate 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 recently described in the INTERVAL study: an expanded SOMAScan panel (3622 protein measurements) identified 1,927 regulatory SNPs for protein levels[3]. This study demonstrated that 14% of aptamers showed non-specific binding; 7% bound more than one protein, and the remaining 7% to 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 colleagues, 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 will 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. Previously, proteomic investigations have been only 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 USA, where the eMERGE cohort is recruited, EHR codes used for billing and costing services. In the UK, the same codes are used by primary care physicians to document medical diagnoses in part to meet incentivised 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 under-diagnosis of rare diseases, and over-reporting of diseases in incentivised 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 MMP-12, which 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, whilst others are missing. ApoE levels have been previously identified as a biomarker for cardiovascular risk by direct measurement with mass spectrometry, the gold standard for specific protein identification and quantification, which avoids reliance on affinity-based techniques[11–13]. On the other hand, other well-known SNP-regulated plasma proteins were not detected. For example, ApoB 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 1,128 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 impacting 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 Randomisation instruments which do not have epitope effects. In particular, changes in electrical 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 pQTLs, 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]*”*

Sources of Funding

A.J. is a British Heart Foundation Clinical Research Training Fellow (FS/16/32/32184). M.M. is a British Heart Foundation Chair Holder (CH/16/3/32406) with British Heart Foundation programme grant support (RG/16/14/32397).

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 [PubMed: 25291143]

[2]. Mosley J, Benson M, Smith JG, Melander O, Ngo D, Shaffer CM, Ferguson J, Matthew H, McCarty C, Chute C, Jarvik G, et al. Probing the virtual proteome to identify novel disease biomarkers. Circulation. 2018 [In press.].

[3]. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, James R, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, et al. Genomic atlas of the human plasma proteome. Nature. 2018; 558:73–79. DOI: 10.1038/s41586-018-0175-2 [PubMed: 29875488]

[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, et al. 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, et al. 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 [PubMed: 21383194]

[7]. Pasea L, Chung SC, Pujades-Rodriguez M, Moayyeri A, Denaxas SC, Fox KAA, Wallentin L, Pocock SJ, Timmis A, Banerjee A, Patel RS, et al. 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 [PubMed: 28329300]

[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 M, 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:e28–e292. 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, et al. 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 [PubMed: 28209220]

[12]. Mayr M, Gerszten R, Kiechl S. Cardiovascular Risk Beyond Low-Density Lipoprotein Cholesterol. J Am Coll Cardiol. 2018; 71(6):633–635. DOI: 10.1016/j.jacc.2017.12.040 [PubMed: 29420959]

[13]. Yin X, Baig F, Haudebourg E, Blankley RT, Gandhi T, Müller S, Reiter L, Hinterwirth H, Pechlaner R, Tsimikas S, Santer P, et al. Plasma Proteomics for Epidemiology: Increasing Throughput With Standard-Flow Rates. Circ Cardiovasc Genet. 2017; 10 pii: 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, et al. Co-regulatory networks of human serum proteins link genetics to disease. Science. 2018; 361(6404):769–773. DOI: 10.1126/science.aaq1327 [PubMed: 30072576]

[15]. Wittgenstein L. Tractatus logico-philosophicus. New York: Routledge Classics; 2001.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

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 non-target human proteins, or to different isoforms of the same human protein.	7% bind to a different protein 7% bind to an isoform[3].
Non-specificity	Aptamers are not specific to human proteins but bind to proteins from other species.	27.6% of aptamers detect proteins in chicken plasma with tenfold 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.	35% of aptamers display altered binding affinity due to SNPs [3].
Batch Effects	Because aptamers do not provide absolute quantification, calibrators are required to ensure a constant measurement across each plate, and normalisation strategies are applied.	Intraplate coefficient of variability can be as high as 3.6%; comparable to the size of some reported genetic effects[5].
Protein Complexes	Plasma proteins can form complexes which 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.