SomaScan Bioinformatics: Normalization, Quality Control, and Assessment of Pre-Analytical Variation

Julián Candia

Translational Gerontology Branch, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA

Email: julian.candia@nih.gov

ABSTRACT

SomaScan is an aptamer-based proteomics assay designed for the simultaneous measurement of thousands of human proteins with a broad range of endogenous concentrations. In its most current version released on November 1, 2023, the 11K SomaScan assay v5.0 is capable of measuring 10,776 human proteins covering major biological processes and disease areas, including cardiology, inflammation, neurology, and oncology. Here, I review bioinformatic approaches to perform normalization, quality control, and variability assessments.

The SomaScan assay

SomaScan[1,2] is a highly multiplexed, aptamer-based assay capable of simultaneously measuring thousands of human proteins broadly ranging from femto- to micro-molar concentrations. This platform relies upon a new generation of protein-capture SOMAmer (*Slow Offrate Modified Aptamer*) reagents[3]. SOMAmers are based on single-stranded, chemically-modified nucleic acids, selected via the so-called SELEX (*Systematic Evolution of Ligands by EXponential enrichment*) process, which is designed to optimize high affinity, slow off-rate, and high specificity to target proteins. These targets extensively cover major molecular functions including receptors, kinases, growth factors, and hormones, and span a diverse collection of secreted, intracellular, and extracellular proteins or domains. In recent years, SomaScan has increasingly been adopted as a powerful tool for biomarker discovery across a wide range of diseases and conditions, as well as to elucidate their biological underpinnings in proteomics and multi-omics studies[4–15].

Concurrently with its wider adoption, SomaScan has expanded its proteome coverage by increasing the number of SOMAmers included in different versions of the assay, from roughly 800 SOMAmers in 2009, to 1,100 in 2012, 1,300 in 2015, 5,000 in 2018, 7,000 in 2020 and the most recent 11,000 protein assay available since late 2023 (Fig. 1). The first independent analysis of SomaScan normalization procedures and their variability was published by our teams at the U.S. National Institutes of Health (NIH) on the 1.1K and 1.3K assays[16], later followed by technical reports from other laboratories[17–21] and a recently updated assessment based on the 7K assay[22].

Table 1 shows the distribution of SOMAmer types in each SomaScan assay version. While most SOMAmers target human proteins, a small fraction of them targets mouse proteins. The remaining SOMAmers are different types of control, including twelve HCE (*Hybridization Control Elution*) SOMAmers used in the hybridization control normalization step (described below), as well as a few non-cleavable SOMAmers, non-biotin SOMAmers, spuriomers (designed as random, non-specific sequence motifs), and legacy SOMAmers targeting proteins from other species. In order to cover a broad range of endogenous concentrations, SOMAmers are binned into different dilution groups, namely 20% (1:5) dilution for proteins typically observed in the femto- to pico-molar range (which comprise about 80% of all human protein SOMAmers in the assay), 0.5% (1:200) dilution for proteins typically present in nano-molar concentrations (slightly below 20% of human protein SOMAmers in the assay), and 0.005% (1:20,000) dilution for proteins in micro-molar concentrations (about $2-3$% of human protein SOMAmers in the assay). Although the proportions of SOMAmers across dilution groups have remained quite stable from version 5K onwards, it should be noticed that version 1.3K and earlier implemented a different scheme based on 40%, 1%, and 0.005% dilution groups. The human plasma or serum volume required is 55 μL per sample.

SOMAmers are uniquely identified by their "SeqId", but the relation between SOMAmers and annotated proteins is not one-to-one. In the 11K assay, for instance, the 10,776 SOMAmers that target human proteins are mapped to 9,609 unique UniProt IDs, 9,550 unique Entrez gene IDs, and 9,604 unique Entrez gene symbols. In the 7K assay, the 7,289 SOMAmers that target human proteins are mapped to 6,383 unique UniProt IDs, 6,378 unique Entrez gene IDs, and 6,383 unique Entrez gene symbols. Among cases where two or more SOMAmers share the same annotated target, the observed RFU correlation

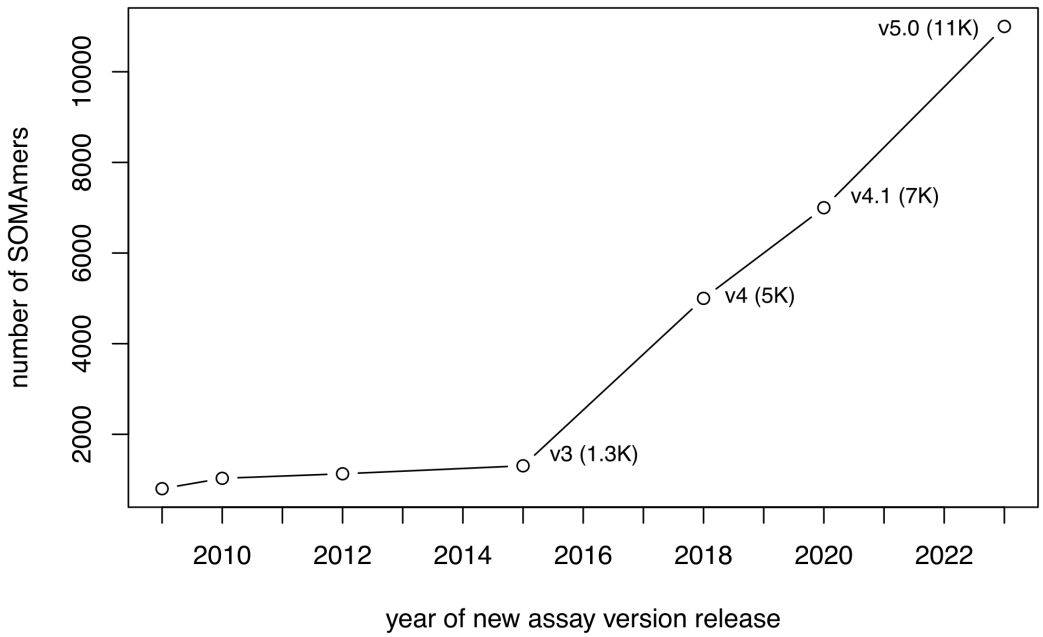

Figure 1. Timeline of SomaScan assay versions. Approximate timeline of assay version release dates showing the growth in the number of SOMAmers in each assay. Since 2015, the assay has been growing at a steady, approximately linear rate.

distribution is bi-modal, with peaks around $r \approx 0$ and $r \approx 1$[22]. While the $r \approx 1$ peak simply indicates the expected redundancy of SOMAmers that bind to the same target, the $r \approx 0$ peak may be due to SOMAmers that bind to different proteoforms annotated under the same protein target name, although they may also be due to artifacts such as cross-reactivity and non-specific binding. All of the SOMAmers in the 7K and 5K SomaScan assays, as well as most of those in the 1.3K version, are included in the newest 11K assay (Fig. 2). Data processing and delivery by SomaLogic is in the `adat` format, which is described in SomaLogic's public GitHub repository (https://github.com/SomaLogic) and has remained consistent across assay versions. The data normalization procedures described below are equally applicable to all assay versions.

The experimental workflow, summarized in Fig. 3, consists of a sequence of steps, namely: (1) SOMAmers are synthesized with a fluorophore, photocleavable linker, and biotin; (2) diluted samples are incubated with dilution-specific SOMAmers bound to streptavidin beads; (3) unbound proteins are washed away, and bound proteins are tagged with biotin; (4) UV light breaks the photocleavable linker, releasing complexes back into solution; (5) non-specific complexes dissociate while specific complexes remain bound; (6) a polyanionic competitor is added to prevent rebinding of non-specific complexes; (7) biotinylated proteins (and bound SOMAmers) are captured on new streptavidin beads; and (8) after SOMAmers are released from the complexes by denaturing the proteins, fluorophores are measured following hybridization to complementary sequences on a microarray chip. The fluorescence intensity detected on the microarray, measured in RFU (*Relative Fluorescence Units*), is assumed to reflect the amount of available epitope in the original sample.

Samples are organized in 96−well plates, each plate consisting of buffer wells (reagent blanks with no sample material

Table 1. Distribution of SOMAmer types

SOMAmer Type vs Assay	11K	7K	5K	1.3K
Human Protein	10776	7289	4979	1302
Mouse Protein	233	233	228	-
Hybridization Control Elution	12	12	12	12
Non-biotin	10	10	10	-
Non-cleavable	4	4	4	-
Spuriomer	20	20	20	-
Other	28	28	31	8

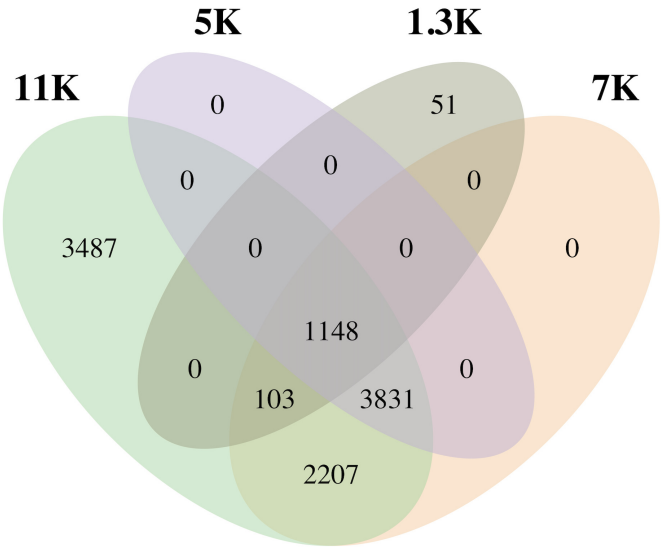

Figure 2. Human protein SOMAmers across SomaScan versions. Venn diagram showing the SOMAmer overlap (based on "SeqId" identifiers) between the 1.3K, 5K, 7K, and 11K SomaScan assays.

added), calibrator and QC samples (which are provided by SomaLogic from pooled healthy donor controls), and the experimental samples of interest. SomaScan users may be interested in adding their own set of control samples to the plate design, which would allow them to bridge across studies independently from SomaLogic-provided controls. On the one hand, SomaLogic's controls may change over time, making it difficult to compare between studies run at different times. On the other hand, owning a set of control samples may allow users to use other omics technologies, including other proteomic assays such as mass spectrometry and Olink[23], for further validation. As a standard practice, our labs at the NIH have implemented the use of a control sample derived from pooled healthy donors, which is run on every plate with 3 to 4 technical replicates per plate.

Data Normalization Procedures

Raw data, as obtained after aggregation from slide-based hybridization microarrays, exhibit intra-plate nuisance variance due to differences in loading volume, leaks, washing conditions, etc, which is then compounded with batch effects across plates. In order to account for intra- and inter-plate variability of buffer, calibrator, QC, and experimental samples, here we consider a sequence of steps, whose main elements are hybridization normalization, median signal normalization, plate-scale normalization, and inter-plate calibration. Each of these steps generates scale factors at different levels: plate-specific, by SOMAmer dilution group, SOMAmer-specific, by sample type, and combinations thereof. Besides removing technical variability, these scale factors can be used as quality control flags at the plate-, sample-, and SOMAmer- levels[8]. A summary of these data normalization procedures, explained below in detail, is shown in Table 2. To keep a consistent mathematical notation in the expressions below, wells are indicated by Latin sub-indices and SOMAmers by Greek sub-indices. Plate index, sample type, and SOMAmer groupings are denoted by super-indices. We use *well* and *sample* interchangeably, although it should be noticed that no sample material is added to buffer wells; also, depending on context, *sample* may refer specifically to the

Table 2. Summary of normalization steps

Step	Name	Abbreviation	Scale Factors
0	Raw	raw	none
1	Hybridization control normalization	hyb	well-specific
2	Median signal normalization on calibrators	hyb.msnCal	calibrator- and dilution-specific
3	Plate-scale normalization	hyb.msnCal.ps	plate-specific
4	Inter-plate calibration	hyb.msnCal.ps.cal	plate- and SOMAmer-specific
5	Median signal normalization on all sample types	hyb.msnCal.ps.cal.msnAll	well-specific (grouped by type)

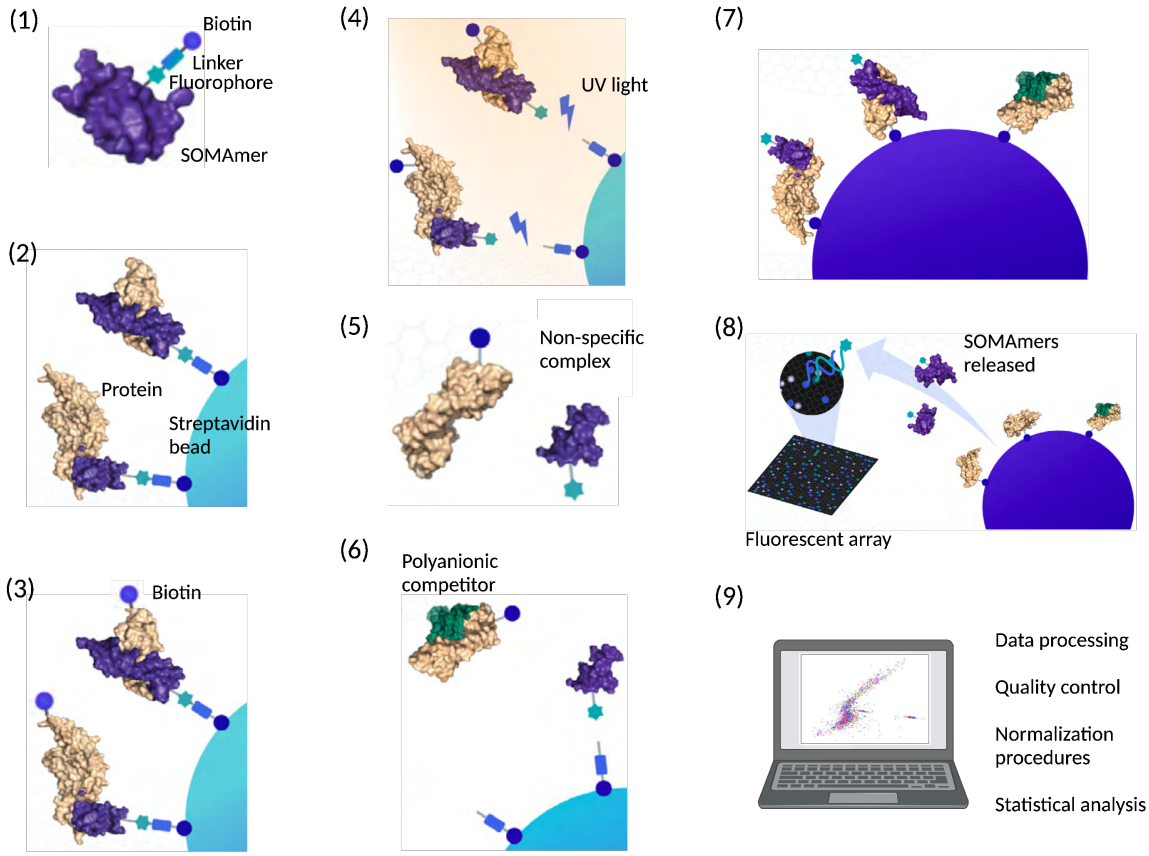

Figure 3. Workflow of the SomaScan assay. (1) SOMAmers are synthesized with a fluorophore, photocleavable linker, and biotin; (2) diluted samples are incubated with dilution-specific SOMAmers bound to streptavidin beads; (3) unbound proteins are washed away, and bound proteins are tagged with biotin; (4) UV light breaks the photocleavable linker, releasing complexes back into solution; (5) non-specific complexes dissociate while specific complexes remain bound; (6) a polyanionic competitor is added to prevent rebinding of non-specific complexes; (7) biotinylated proteins (and bound SOMAmers) are captured on new streptavidin beads; and (8) after SOMAmers are released from the complexes by denaturing the proteins, fluorophores are measured following hybridization to complementary sequences on a microarray chip. Upon completion of all experimental steps, (9) the bioinformatic analysis proceeds. Adapted from SomaLogic's Technical Note SL00000572 and created with BioRender.

experimental samples of interest, thus excluding control sample types such as buffer, calibrator, and QC.

As we describe each normalization step, we also include R code chunks that implement each of those steps. We assume that the raw `adat` file is available to the user and that it has been split into three different plain-text files, namely: (i) "samples.txt", with headers and one sample per row, providing metadata information on each well in the study; (ii) "somamers.txt", with headers and one SOMAmer per row, providing metadata information on each SOMAmer in the assay; and (iii) "RFU.raw.txt", without headers, samples as rows and SOMAmers as columns, providing the measured raw RFU values for each well and SOMAmer in the study. The ordering of samples and SOMAmers in "RFU.raw.txt" must be consistent with that of "samples.txt" and "somamers.txt", respectively. Files in the `adat` format can be opened and manipulated programmatically via custom scripts, using the `SomaDataIO` R package available from SomaLogic's public GitHub repository (https://github.com/SomaLogic), or using previously developed open-source tools[24,25]; however, it is often simple enough to open the `adat` file in Excel or similar software, then copy-paste the three separate regions corresponding to sample metadata, SOMAmer metadata (that needs to be transposed in order to display SOMAmers row-wise) and RFU matrix, and save them separately as tab-delimited plain-text files. Fig. 4 shows an example of the input data files required for this normalization pipeline.

The following code chunk reads the input datasets and sets up the variables needed to implement each of the normalization steps described below. Notice that, for the sake of clarity, we avoid using variable names that overlap with names of core R functions (e.g. `sample`). The code presented here was written with an emphasis on readability but not optimized for computing

a

PlateId	PlatePosition	SampleId	SampleType	SampleMatrix	Barcode	ExtIdentifier
P0031168	A1	5302-15	Sample	EDTA plasma	S1545380	EXID40000006690911
P0031168	A10	4875-1	Sample	EDTA plasma	S1545524	EXID40000006690909
P0031168	A11	824-29	Sample	EDTA plasma	S1545540	EXID40000006690925
P0031168	A12	4922-2	Sample	EDTA plasma	S1545442	EXID40000006690906
P0031168	A2	4913-2	Sample	EDTA plasma	S1545396	EXID40000006690763
P0031168	A3	1550-11	Sample	EDTA plasma	S1545412	EXID40000006690800
P0031168	A4	1768-4	Sample	EDTA plasma	S1545428	EXID40000006690756
P0031168	A5	7828-1	Sample	EDTA plasma	S1545444	EXID40000006690765
P0031168	A6	5619-11	Sample	EDTA plasma	S1545460	EXID40000006690936
P0031168	A7	7862-1	Sample	EDTA plasma	S1545476	EXID40000006690753

b

SeqId	SomaId	Target	UniProt	EntrezGeneID	EntrezGeneSymbol	Organism	Type	Dilution
10000-28	SL019233	CRBB2	P43320	1415	CRYBB2	Human	Protein	20
10001-7	SL002564	c-Raf	P04049	5894	RAF1	Human	Protein	20
10003-15	SL019245	ZNF41	P51814	7592	ZNF41	Human	Protein	0.5
10006-25	SL019228	ELK1	P19419	2002	ELK1	Human	Protein	20
10008-43	SL019234	GUC1A	P43080	2978	GUCA1A	Human	Protein	20
10010-10	SL014943	BECN1	Q14457	8678	BECN1	Human	Protein	20
10011-65	SL019246	OCRL	Q01968	4952	OCRL	Human	Protein	20
10012-5	SL014669	SPDEF	O95238	25803	SPDEF	Human	Protein	20
10013-34	SL025418	Fc_MOUSE	Q99LC4			Mouse	Protein	20
10014-31	SL007803	SLUG	O43623	6591	SNAI2	Human	Protein	20

c

748.8	325.4	217.1	572	720.7	357.6	2323.9	1382.1	421.7	859.7
577.5	385.4	325.7	634.8	731	413.9	2847.4	2004	484.7	936.9
526.8	318.6	203	581.3	618	317.6	2510.5	1370.7	412.9	796.9
508.4	422.7	213.1	641.2	688.2	340.4	2366	1317.6	393.7	850.5
450.5	290.5	202.2	580.2	2294.6	277.9	2108	1210.6	356.6	694.6
621.3	456	204.3	663.7	599	814.1	2486	1362.2	445.9	861.5
658.4	508.4	226.7	821.2	681	440.9	3236.4	1772.6	505.3	1013.5
571.8	385.6	221.3	612.3	633.5	373.1	2849.8	1890	446.4	852.3
482	312.1	191	575.3	565.7	334	2792	1286.4	390.7	776.7
5986.7	300.4	206.7	538.1	651.4	307.3	2293.8	1271.2	407.5	762.7

Figure 4. Input data files for the normalization pipeline. a "samples.txt", with headers and one sample per row, provides metadata information on each well in the study (including buffer, QC, calibrator, and experimental samples). **b** "somamers.txt", with headers and one SOMAmer per row, provides metadata information on each SOMAmer in the assay (including controls such as Hybridization Control Elution (HCE) SOMAmers). **c** "RFU.raw.txt", without headers, samples as rows and SOMAmers as columns, provides the measured raw RFU values for each well and SOMAmer in the study (in the same order as the sample and SOMAmer metadata files).

performance. Since `adat` files are much smaller than typical datasets generated by other omics and bioinformatic tasks, further optimization was not deemed necessary.

```
norm = c("raw","hyb","hyb.msnCal","hyb.msnCal.ps","hyb.msnCal.ps.cal","hyb.msnCal.ps.cal.msnAll")
n_norm = length(norm)
RFU = vector("list",n_norm)

sampl = as.matrix(read.table("samples.txt",header=T,quote="",comment.char="",sep="\t"))
somamers= as.matrix(read.table("somamers.txt",header=T,quote="",comment.char="",sep="\t"))
RFU[[1]] = as.matrix(read.table("RFU.raw.txt",header=F,sep="\t"))

n_sampl = nrow(sampl)
n_somamer = nrow(somamer)
plate = unique(sampl[,"PlateId"])
n_plate = length(plate)

dil = c(0.005,0.5,20)
dil_lab = c("0_005","0_5","20")
n_dil = length(dil)
```

1. Hybridization control normalization (hyb)

Hybridization control normalization is designed to adjust for nuisance variance on the basis of individual wells. Each well contains $n_{HCE} = 12$ HCE (*Hybridization Control Elution*) SOMAmers at different concentrations spanning more than 3 orders of magnitude. By comparing each observed HCE probe to its corresponding reference value and then calculating the median over all HCE probes, we obtain the scale factor for the i-th well, i.e.

$$SF_i = \text{median}\left\{\frac{RFU_{\alpha}^{HCE,ref}}{RFU_{i\alpha}^{HCE,obs}}\right\}_{\alpha=1,\ldots,n_{HCE}} . \tag{1}$$

Notice that this normalization step is performed independently for each well; once the scale factor is determined, all SOMAmer RFUs in the well are multiplied by the same scale factor. Instead of using an external reference, a plate-specific internal reference can be determined by the median across the $n_s = 96$ wells in the plate, i.e.

$$RFU_{\alpha}^{HCE,ref} \equiv \text{median}\left\{RFU_{i\alpha}^{HCE,obs}\right\}_{i=1,\ldots,n_s} . \tag{2}$$

The following code chunk implements the hybridization control normalization step:

```
sel_HCE = somamer[,"Type"]=="Hybridization Control Elution"
n_HCE  = sum(sel_HCE)
RFU_HCE = RFU[[1]][,sel_HCE]

RFU[[2]] = RFU[[1]]
SF_hyb = rep(NA,n_sampl)

HCE_ratio = matrix(rep(NA,n_sampl*n_HCE),ncol=n_HCE)
for (i_plate in 1:n_plate) {
    sampl_sel = sampl[,"PlateId"]==plate[i_plate]
    hyb_intra_ref = apply(RFU_HCE[sampl_sel,],2,median)
    HCE_ratio_plate = HCE_ratio[sampl_sel,]
    SF_hyb_plate = SF_hyb[sampl_sel]
    RFU_plate = RFU[[2]][sampl_sel,]
    for (i_sampl in 1:sum(sampl_sel)) {
        HCE_ratio_plate[i_sampl,] = hyb_intra_ref/RFU_plate[i_sampl,sel_HCE]
        SF_hyb_plate[i_sampl] = median(HCE_ratio_plate[i_sampl,])
        RFU_plate[i_sampl,] = SF_hyb_plate[i_sampl]*RFU_plate[i_sampl,]
    }
    HCE_ratio[sampl_sel,] = HCE_ratio_plate
    SF_hyb[sampl_sel] = SF_hyb_plate
    RFU[[2]][sampl_sel,] = RFU_plate
}
```

2. Median signal normalization on calibrators (hyb.msnCal)

Median signal normalization is an intra-plate normalization procedure performed within wells of the same sample class (i.e. separately for buffer, QC, calibrator, and experimental samples) and within SOMAmers of the same dilution group. It is intended to remove sample-to-sample differences in total RFU brightness that may be due to differences in overall protein concentration, pipetting variation, variation in reagent concentrations, assay timing, and other sources of variability within a group of otherwise comparable samples. Since RFU brightness differs significantly across SOMAmers, median signal normalization proceeds in two steps. First, the median RFU of each SOMAmer is determined (across all samples of the same sample type) and sample RFUs are divided by it. The ratio corresponding to the i-th sample and α-th SOMAmer is thus given by

$$r_{i\alpha}^{gd} = RFU_{i\alpha}/\text{median}\left\{RFU_{j\alpha}\right\}_{j=1,\ldots,n_s^g} , \tag{3}$$

where indices g and d denote sample type and SOMAmer dilution groupings, respectively. Then, the scale factor associated with the i-th sample is determined as the inverse of the median ratio for that sample across all SOMAmers in the dilution group:

$$SF_i^{gd} = 1/\text{median}\left\{r_{i\alpha}^{gd}\right\}_{\alpha=1,\ldots,n_{SOMAmer}^d} . \tag{4}$$

To median-normalize the i-th sample, then, all its SOMAmer RFUs in the same dilution group are multiplied by this scale factor. This procedure is illustrated by Fig. 5. As discussed in our previous work[16], performing median signal normalization on experimental samples *before* inter-plate calibration presents the risk of enhancing plate-to-plate differences. Thus, in this step, we restrict median signal normalization to calibrators only.

The following code chunk implements the median signal normalization step on calibrators:

```
RFU[[3]] = RFU[[2]]
SF_medNormInt = matrix(rep(1,n_sampl*n_dil),ncol=n_dil)

sampl_type = c("Calibrator")
n_sampl_type = length(sampl_type)

for (i_plate in 1:n_plate) {
    sel1 = sampl[,"PlateId"]==plate[i_plate]
    for (i_sampl_type in 1:n_sampl_type) {
        sel2 = sampl[,"SampleType"]==sampl_type[i_sampl_type]
        indx_sampl = which(sel1&sel2)
        for (i_dil in 1:n_dil) {
            sel_somamer = as.numeric(somamer[,"Dilution"])==dil[i_dil]
            dat = RFU[[3]][indx_sampl,sel_somamer,drop=F]
            dat2 = dat
            for (i in 1:ncol(dat2)) {
                dat2[,i] = dat2[,i]/median(dat2[,i])
            }
            for (i in 1:nrow(dat)) {
                SF_medNormInt[indx_sampl[i],i_dil] = 1/median(dat2[i,])
                dat[i,] = dat[i,]*SF_medNormInt[indx_sampl[i],i_dil]
            }
            RFU[[3]][indx_sampl,sel_somamer] = dat
        }
    }
}
```

3. Plate-scale normalization (hyb.msnCal.ps)

Plate-scale normalization aims to control for variance in total signal intensity from plate to plate. No protein spikes are added to the calibrator; the procedure solely relies on the endogenous levels of each protein within the set of calibrator replicates.

For the α-th SOMAmer on the p-th plate,

$$SF_{\alpha}^{p} = RFU_{\alpha}^{Cal,ref} / \text{median} \left\{ RFU_{i\alpha}^{Cal,obs} \right\}_{i=1,\ldots,n_{Cal}}^{p} . \tag{5}$$

Calibration scale factors may be pinned to an external reference, but here we utilize an internal reference determined by the median across all calibrators on all n_p plates, i.e.

$$RFU_{\alpha}^{Cal,ref} \equiv \text{median} \left\{ RFU_{i\alpha}^{Cal,obs} \right\}_{i=1,\ldots,n_{Cal}}^{p=1,\ldots,n_p} . \tag{6}$$

In order to correct the overall brightness level of the p-th plate, we calculate the plate-scale scale factor as the median of SF_{α}^{p} across all SOMAmers, i.e.

$$SF^{p} = \text{median} \left\{ SF_{\alpha}^{p} \right\}_{\alpha=1,\ldots,n_{SOMAmer}} . \tag{7}$$

For all wells on the p-th plate and all SOMAmers, RFUs are multiplied by the plate-scale factor SF^{p}.

The following code chunk implements the plate-scale normalization step:

```
RFU[[4]] = RFU[[3]]

SF = matrix(rep(NA,n_plate*n_somamer),ncol=n_somamer)
```

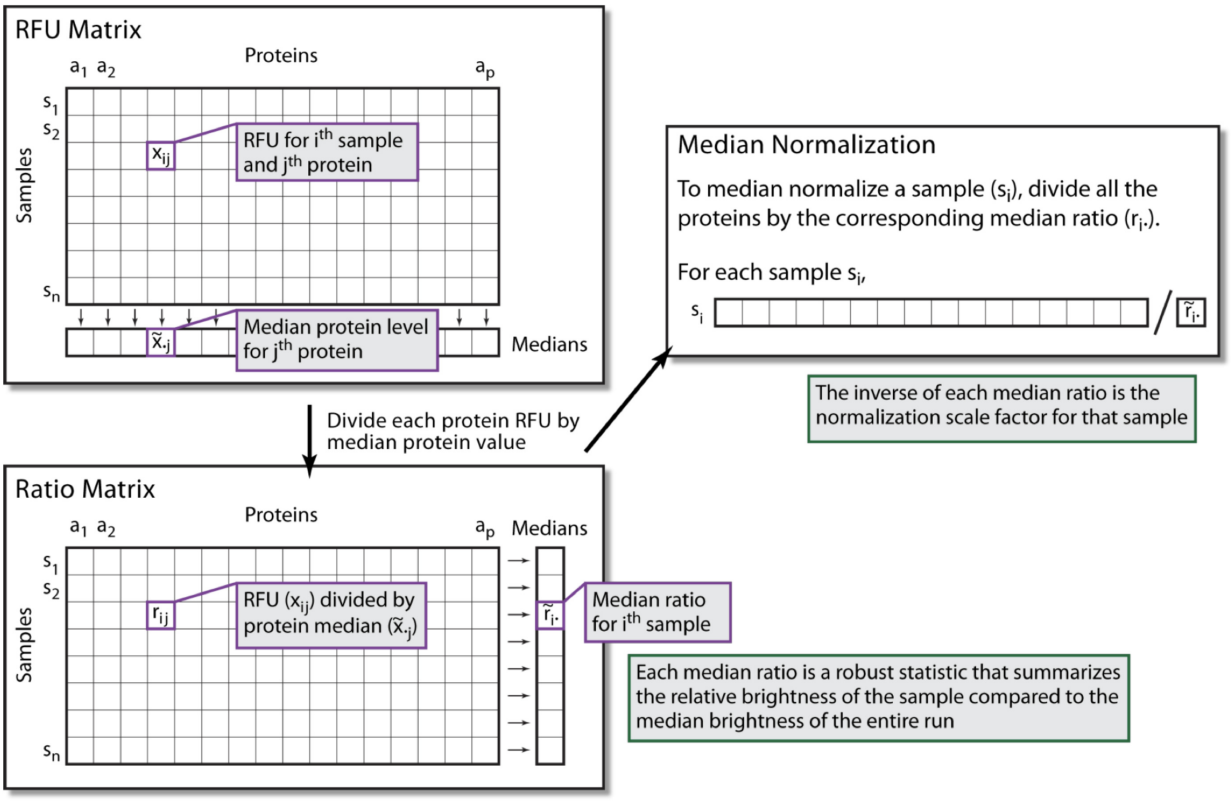

Figure 5. Median signal normalization. An RFU sub-matrix is defined by all samples of the same class (i.e. separately for buffer, QC, calibrator, and experimental samples) and all SOMAmers within the same dilution group. For each SOMAmer, the median RFU level is calculated across all samples (top left). By dividing each RFU measurement by the corresponding SOMAmer median, we obtain the ratio matrix; then, for each sample, a median ratio is calculated (bottom left). The median normalization scale factor associated to each sample is the inverse of the median ratio. To median-normalize each sample, all SOMAmers within the target dilution group are multiplied by its corresponding scale factor. Adapted from image courtesy of Darryl Perry (SomaLogic).

```
sel_cal = sampl[,"SampleType"]=="Calibrator"
cal_interplate_ref = rep(NA,n_somamer)
for (i_somamer in 1:n_somamer) {
    cal_interplate_ref[i_somamer] = median(RFU[[4]][sel_cal,i_somamer])
    for (i_plate in 1:n_plate) {
        sel_plate = sampl[,"PlateId"]==plate[i_plate]
        cal_intraplate_ref = median(RFU[[4]][sel_cal&sel_plate,i_somamer])
        SF[i_plate,i_somamer] = cal_interplate_ref[i_somamer]/cal_intraplate_ref
    }
}
SF_plateScale = apply(SF,1,median)

for (i_plate in 1:n_plate) {
    sel_plate = sampl[,"PlateId"]==plate[i_plate]
    RFU[[4]][sel_plate,] = RFU[[4]][sel_plate,]*SF_plateScale[i_plate]
}
```

4. Inter-plate calibration (hyb.msnCal.ps.cal)

Following plate-scale normalization, we recalculate SOMAmer- and plate-specific scale factors via Eqs. (5)-(6). Separately for each SOMAmer and plate, all wells on the plate are corrected by the recalculated SF_{α}^{p}.

The following code chunk implements the inter-plate calibration step:

```
RFU[[5]] = RFU[[4]]

SF_cal = matrix(rep(NA,n_plate*n_somamer),ncol=n_somamer)
sel_cal = sampl[,"SampleType"]=="Calibrator"
cal_interplate_ref_N = rep(NA,n_somamer) # new interplate reference
for (i_somamer in 1:n_somamer) {
    cal_interplate_ref_N[i_somamer] = median(RFU[[5]][sel_cal,i_somamer])
    for (i_plate in 1:n_plate) {
        sel_plate = sampl[,"PlateId"]==plate[i_plate]
        cal_intraplate_ref = median(RFU[[5]][sel_cal&sel_plate,i_somamer])
        SF_cal[i_plate,i_somamer] = cal_interplate_ref_N[i_somamer]/cal_intraplate_ref
        RFU[[5]][sel_plate,i_somamer] = RFU[[5]][sel_plate,i_somamer]*SF_cal[i_plate,i_somamer]
    }
}
```

5. Median signal normalization on all sample types (hyb.msnCal.ps.cal.msnAll)

At this stage, after correcting plate-to-plate variability to the fullest extent possible, median signal normalization (described in step 2 and Fig. 5) can be performed separately on each sample type. This step yields the final, fully-normalized dataset. The following code chunk implements the median signal normalization step on all sample types:

```
RFU[[6]] = RFU[[5]]
SF_medNormFull = matrix(rep(1,n_sampl*n_dil),ncol=n_dil)

sampl_type = c("QC","Sample","Buffer","Calibrator")
n_sampl_type = length(sampl_type)

for (i_sampl_type in 1:n_sampl_type) {
    indx_sampl = which(sampl[,"SampleType"]==sampl_type[i_sampl_type])
    for (i_dil in 1:n_dil) {
        sel_somamer = as.numeric(somamer[,"Dilution"])==dil[i_dil]
        dat = RFU[[6]][indx_sampl,sel_somamer,drop=F]
        dat2 = dat
        for (i in 1:ncol(dat2)) {
            dat2[,i] = dat2[,i]/median(dat2[,i])
        }
        for (i in 1:nrow(dat)) {
            SF_medNormFull[indx_sampl[i],i_dil] = 1/median(dat2[i,])
            dat[i,] = dat[i,]*SF_medNormFull[indx_sampl[i],i_dil]
        }
        RFU[[6]][indx_sampl,sel_somamer] = dat
    }
}
```

Finally, the following code chunk saves all sample-, SOMAmer-, and plate-based metadata (including the scale factors used in the normalization procedure) along with the RFU matrices at each normalization step:

```
header = c(colnames(sampl),"SF_hyb",paste0("SF_msnCal_d",dil_lab),paste0("SF_msnAll_d",dil_lab))
output = rbind(header,cbind(sampl,SF_hyb,SF_medNormInt,SF_medNormFull))
write(t(output),ncol=ncol(output),file="samples_SF.txt",sep="\t")

header = c(colnames(somamer),"Cal_Interplate_Ref_Pass1","Cal_Interplate_Ref_Pass2")
output = rbind(header,cbind(somamer,cal_interplate_ref,cal_interplate_ref_N))
```

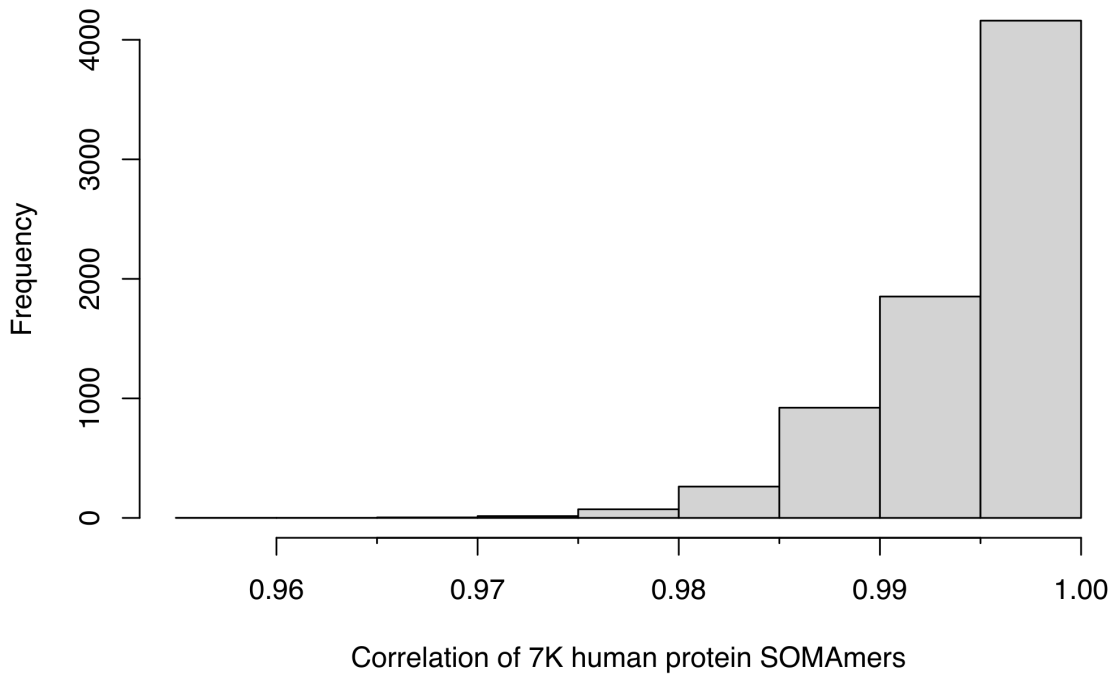

Figure 6. Concordance between normalization approaches. Distribution of Spearman's correlation estimates for the 7,289 human protein SOMAmers in the 7K plasma SomaScan assay, calculated over nearly 1,800 human donor samples from the BLSA[22]. For each SOMAmer, the correlation was calculated between the fully normalized RFU values from SomaLogic's pipeline using external references and the full normalization described here using internal references.

```
write(t(output),ncol=ncol(output),file="somamers_SF.txt",sep="\t")

header = c("Plate","SF_plateScale",paste0("SF_cal_",somamer[,"SeqId"]))
output = rbind(header,cbind(plate,SF_plateScale,SF_cal))
write(t(output),ncol=ncol(output),file="plates_SF.txt",sep="\t")

for (i_norm in 2:n_norm) {
    outfile = paste0("RFU.",norm[i_norm],".txt")
    write(t(RFU[[i_norm]]),ncol=n_somamer,file=outfile,sep="\t")
}
```

It should be pointed out that SomaLogic currently delivers normalized output files (as plain text files of extension `adat`) that follow similar steps as those described here, but using external references[8]. These references may be outdated (because they are specific to the control samples used), not necessarily representative of the target samples of interest (because they utilize a fixed pool of healthy human control samples for the last normalization step) and are not delivered with the `adat` files provided to the customer (therefore precluding any attempt of independent analysis). In contrast, the process described here relies solely on internal references derived from the measured data. In our study of nearly 1,800 experimental samples from the Baltimore Longitudinal Study on Aging (BLSA), we found that fully normalized datasets using internal versus external references were highly concordant. For each human protein SOMAmer in the plasma 7K assay, we calculated the Spearman's correlation between the fully normalized RFU values from the `adat` file provided by SomaLogic using external references (a file designated with the "hybNorm.medNormInt.plateScale.calibrate.anmlQC.qcCheck.anmlSMP" suffix) and the full normalization described here using internal references ("hyb.msnCal.ps.cal.msnAll"). The distribution of correlation estimates over all 7,289 human protein SOMAmers is shown in Fig. 6; the distribution median is $r = 0.996$. It should be added that, due to Spearman's correlation invariance, these results remain valid under any type of monotonic (e.g. logarithmic) RFU transformations.

SomaLogic's standard data quality reports utilize scale factors from the normalization procedure to flag samples, SOMAmers, and plates that do not pass pre-established acceptance criteria. Additional standard approaches that SomaScan users may implement are dimensional reduction techniques such as Principal Component Analysis (PCA), which serves both as a quality

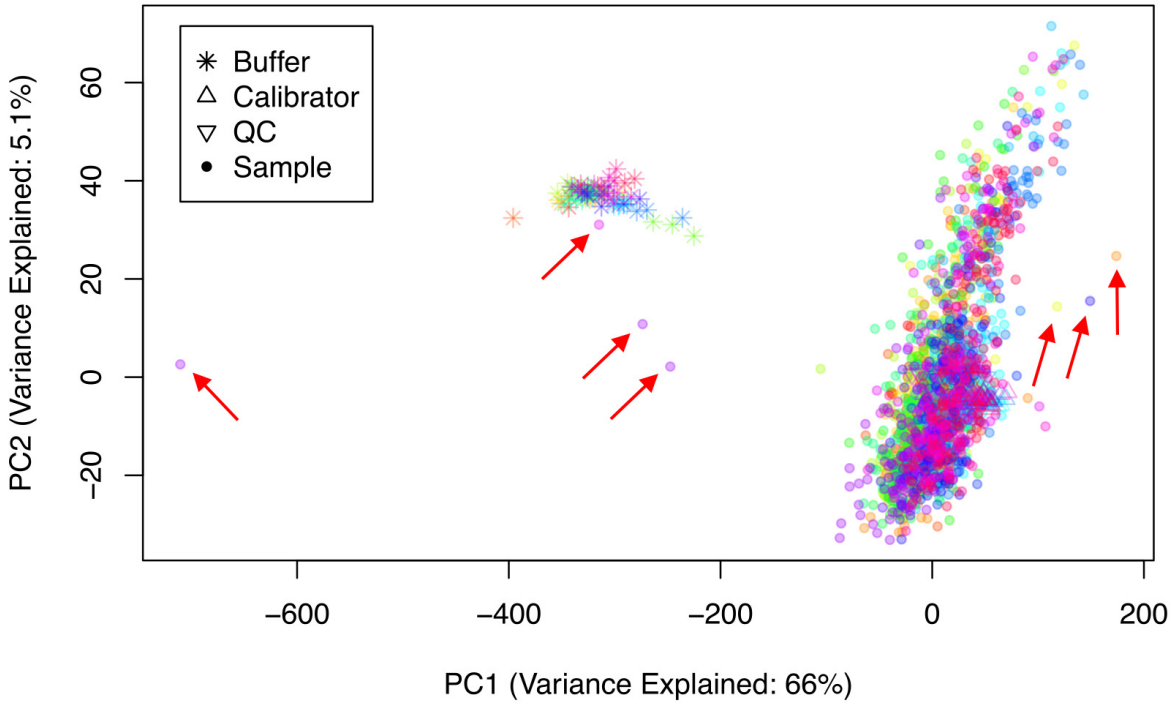

Figure 7. PCA as quality control and data exploration tool. PCA performed on 2,050 raw (not normalized) samples, including 68 buffers, 110 calibrators, 66 QC and 2,806 BLSA samples distributed across 22 plates. Each plate is shown by a different color. The red arrows show seven samples flagged due to their out-of-range normalization scale factors.

control and exploration tool to gain insight on the data. Fig. 7 shows PCA performed on 2,050 raw (not normalized) samples, including 68 buffers, 110 calibrators, 66 QC and 2,806 human donor samples from the BLSA[22]. Seven samples, flagged in SomaLogic's data quality report due to their out-of-range normalization scale factors, are shown by red arrows. In this case, PCA helps to confirm that those samples indeed appear as outliers. Four of them, in fact, appear closer to the reagent-only buffer wells than the large cluster formed by calibrator, QC, and the remaining biological samples. Since this study spans 22 plates, shown by different colors, PCA may provide visual hints of plate effects; inspecting and comparing PCA plots with different normalizations provides more clues to explore whether those potential plate effects were effectively removed by the normalization process[22]. It should be noticed that other bioinformatic tools, developed to assess and remove batch effects for other omics, may be useful in the context of SomaScan data processing, for instance: ComBat[26], implemented in the sva R package[27], a very well established method for batch correction in microarray and RNA-seq data; guided PCA[28]; and multi-MA normalization[29], among others. This important topic certainly deserves further investigation.

Pre-Analytical Variation (PAV) SomaSignal Tests (SSTs)

Pre-analytical variation (PAV) due to sample collection, handling, and storage is known to affect many analyses in molecular biology. By implementing data modeling techniques similar to those previously developed to find SomaScan signatures associated with clinical phenotypes[8], SomaLogic has developed a novel set of so-called SomaSignal Tests (SSTs) to assess pre-analytical variation due to different sample processing factors, including fed-fasted time, number of freeze-thaw cycles, time-to-decant, time-to-spin, and time-to-freeze. While it is not always possible to control sample collection, processing, storage, and handling to ensure consistency, PAV-SSTs were designed to quantitatively estimate whether and to what extent samples may have been impacted by these nuisance factors.

In previous work[22], we implemented a framework to assess technical variability based on measurements performed on duplicate biological inter-plate pairs from samples obtained from $n_{dupl} = 102$ human participants in the BLSA. Panels (a)-(e) of Fig. 8 show, for each PAV metric, the magnitude of the difference between estimates from each duplicate pair, $\Delta x_i = |x_{i,1} - x_{i,2}|$, as a function of their average, $\bar{x}_i = (x_{i,1} + x_{i,2})/2$. Next, we will leverage this framework to generate an independent assessment of the robustness of SomaLogic's PAV estimates. Let us first recall that, based on the scaled relative differences, defined as $D_i = (\Delta x_i / \bar{x}_i)/\sqrt{2}$, different variability estimates may be built, namely: (i) the Root-Mean-Squared Variation (RMSV) metric,

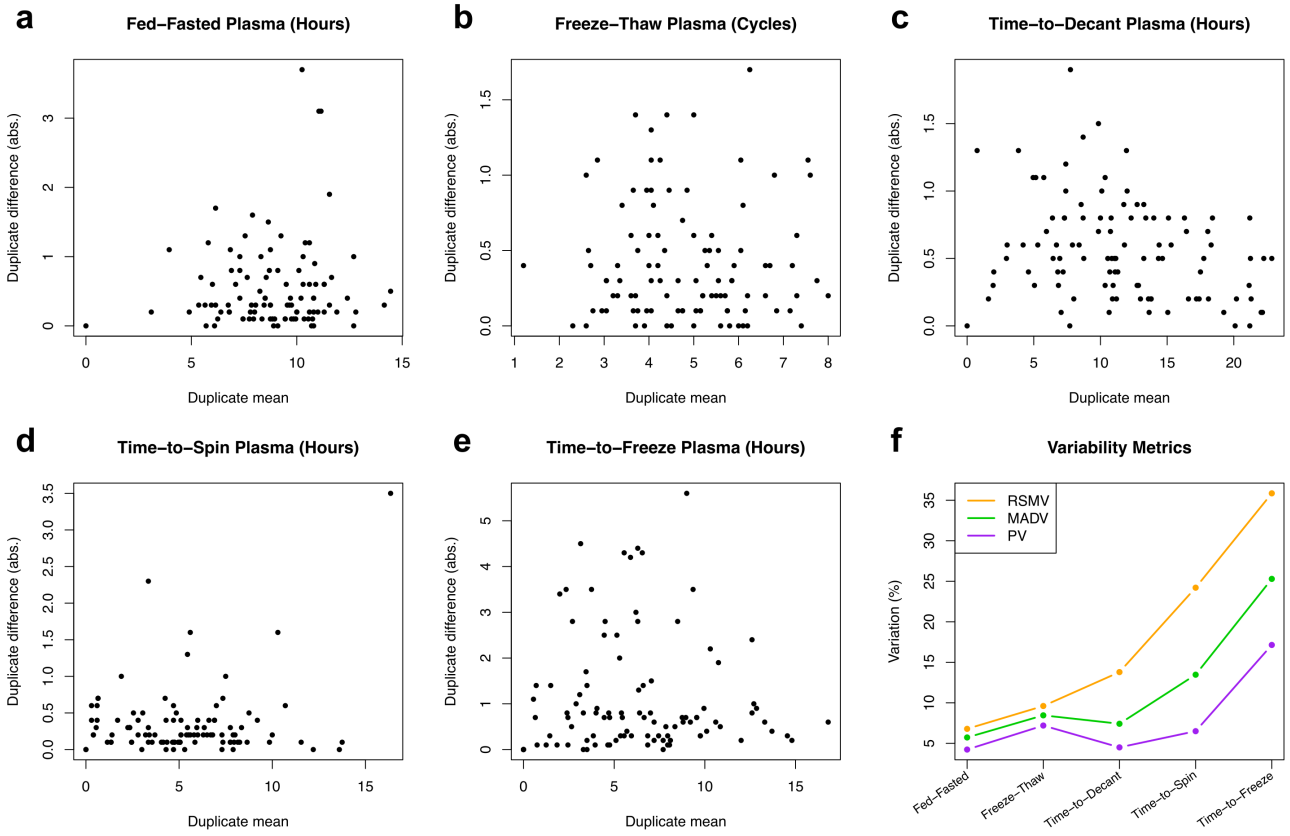

Figure 8. Assessment of the robustness of PAV-SST estimates using BLSA technical replicates. a-e Duplicate difference vs duplicate mean for different PAV-SST estimates, as indicated. Each circle represents one pair of technical duplicates from the same human donor sample. **f** Variation of PAV-SST estimates determined by different metrics, namely: Root-Mean-Squared Variation (RMSV), Mean Absolute Difference Variation (MADV), and Percentile Variation (PV).

defined as

$$RMSV = \sqrt{\frac{1}{n_{dupl}} \sum_{i=1}^{n_{dupl}} D_i^2} \times 100\% \ , \tag{8}$$

(ii) the Mean Absolute Difference Variation (MADV) metric, defined as

$$MADV = \sqrt{\frac{\pi}{2}} \frac{1}{n_{dupl}} \sum_{i=1}^{n_{dupl}} |D_i| \times 100\% \ , \tag{9}$$

and (iii) the Percentile Variation (PV) metric, defined as

$$PV = \frac{1}{2} \left(P_{84}\{D_i\}_{i=1,\ldots,n_{dupl}} - P_{16}\{D_i\}_{i=1,\ldots,n_{dupl}} \right) \times 100\% \ , \tag{10}$$

where P_{84} and P_{16} are the 84^{th} and 16^{th} percentiles in the distribution of scaled relative differences. If the scaled relative differences are normally distributed, these variability estimates are equivalent; the magnitude of the differences between these estimates is indicative of the magnitude and severity of deviations from normality. Armed with these various definitions, Panel Fig. 8(f) shows the percent variation across PAV estimates. As expected by theoretical considerations, the PV metric is the least sensitive to distribution tails and therefore yields the lowest variation. We observe that, with the exception of the time-to-freeze estimate, which shows a PV variation of about 15%, the PV variation for the remaining four PAV metrics lies approximately within the $5-7\%$ range.

Summary and Conclusions

With the addition of the latest v5.0 (11K) version released on November 1, 2023, SomaScan galvanized its role as one of the leading technologies in high-throughput human proteomics. Going back to the v3 (1.3K) version about eight years prior, the assay has shown a steady, approximately linear growth in protein coverage. Several independent teams, including ours at the NIH, have explored the assay's sensitivity and variability[16–22]. Furthermore, multiple studies have investigated SomaScan's specificity, cross-reactivity, and orthogonal assay reproducibility[4,5,30–32]. Less attention, however, has been dedicated to explore normalization procedures. Typically, SomaLogic provides one or several data files in `adat` format, which are normalized using their internal analysis pipeline. This pipeline, however, has historically remained under SomaLogic's control and operated to a large extent as a black box, providing SomaScan users only with descriptive explanations of the steps involved in the process. SomaLogic's normalization procedures, moreover, have evolved overtime, switching from internal references to external references, altering the order in which median normalization was implemented, and splitting the interplate calibration step into plate-level rescaling (referred to as plate-scale normalization) followed by SOMAmer- and plate-specific normalization. To the best of our knowledge, no software version control of these changes have been shared with SomaScan users, making it difficult the task to compare or integrate studies run years apart, perhaps even using different versions of the assay.

As an early contribution to the nascent field of SomaScan bioinformatics, our NIH labs have undertaken the task to reconstruct the steps involved in a normalization procedure that, without the use of SomaLogic's proprietary external references, is capable of reproducing normalized RFU values concordant with those from SomaLogic's pipeline[16,22]. The latest version of our normalization pipeline is described here in detail, including its R code implementation, and available from a public repository. The ability to run independent normalization analyses is important for various reasons. On the one hand, the set of external references (based on a fixed pool of healthy human control samples) may be inappropriate to study individuals and populations with large deviations from a healthy plasma proteome. As noted by Lopez-Silva et al[33] in their study of chronic kidney disease, SomaLogic's normalization could potentially attenuate the strength of associations by reducing more extreme values if they are associated with a clinical outcome, perhaps contributing to the attenuated prognostic associations they observed with SomaScan. Pietzner et al[20] report that using SomaScan data without a normalization step applied to correct for unwanted technical variation and to make data comparable across cohorts, a higher median correlation with results from Olink was observed, along with substantial differences in the association with various phenotypic characteristics. On the other hand, users may be interested in using their own set of control samples to bridge across studies, thereby needing to calibrate different studies using those controls. More generally, users may be interested in adapting the normalization process according to their studies' characteristics and objectives (for instance, splitting the median normalization procedure into different sample subclasses identified by clinical phenotype). Therefore, we believe that presenting this independent normalization pipeline empowers SomaScan end users to tailor the normalization procedures to their own individual needs.

Following our discussion of normalization approaches, we briefly described quality control procedures based on combining flags from outlier normalization scale factors (such as those provided in SomaLogic's standard data quality reports) with Principal Component Analysis (PCA), a linear dimensional reduction technique. Alternatively, non-linear dimensional reduction methods such as Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE) could be used for the purpose of quality control and data exploration. In addition, we suggested that bioinformatic tools developed to assess and remove batch effects in data from other omics, such as ComBat[26], Surrogate Variable Analysis (SVA)[27], guided PCA[28], and multi-MA normalization[29], may be useful techniques that deserve further investigation.

Finally, we investigated SomaLogic's SomaSignal Test (SST) models, which were recently developed to estimate a variety of Pre-Analytical Variation (PAV) metrics, including fed-fasted time, number of freeze-thaw cycles, time-to-decant, time-to-spin, and time-to-freeze. Leveraging technical duplicates from our study of human donors from the Baltimore Longitudinal Study on Aging (BLSA), we performed an independent evaluation of the statistical variation of the PAV estimates. Within the limited scope of our assessment, we conclude that PAV-SST estimates appear to be robust (percentile variation in the $5-7\%$ range, with the exception of time-to-freeze, which is about 15%) and may be used during biomarker evaluations to exclude samples due to nuisance effects related to sample collection and processing, and/or to incorporate these estimates as model covariates to control the downstream impact of those unwanted effects.

Data and Software Availability

Anonymized datasets and R source code for the normalization pipeline described here are available on the Open Science Framework repository, osf.io/srgef. DOI 10.17605/OSF.IO/SRGEF.

Acknowledgements

This work was supported entirely by the Intramural Research Program of the National Institute on Aging (NIA).

Competing interests

The author declares no competing interests.

References

1. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE* **5**, e15004 (2010).
2. Schneider, D. J. *et al.* Chapter 8 - somamer reagents and the somascan platform: Chemically modified aptamers and their applications in therapeutics, diagnostics, and proteomics. In Giangrande, P. H., de Franciscis, V. & Rossi, J. J. (eds.) *RNA Therapeutics*, 171–260 (Academic Press, 2022).
3. Rohloff, J. C. *et al.* Nucleic acid ligands with protein-like side chains: Modified aptamers and their use as diagnostic and therapeutic agents. *Mol. Ther. Nucleic Acids* **3**, e201 (2014).
4. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Sci.* **361**, 769–773 (2018).
5. Sun, B. *et al.* Genomic atlas of the human plasma proteome. *Nat.* **558**, 73 (2018).
6. Tanaka, T. *et al.* Plasma proteomic signature of age in healthy humans. *Aging Cell* **17**, e12799 (2018).
7. Lehallier, B. *et al.* Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Medicine* **25**, 1843–1850 (2019).
8. Williams, S. *et al.* Plasma protein patterns as comprehensive indicators of health. *Nat. Medicine* **25**, 1851–1857 (2019).
9. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Sci.* **374**, 839 (2021).
10. Roberts, J. *et al.* A brain proteomic signature of incipient alzheimer's disease in young apoe $\varepsilon 4$ carriers identifies novel drug targets. *Sci Adv* **7**, eabi8178 (2021).
11. Walker, K. *et al.* Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. *Nat. Aging* **1**, 473–489 (2021).
12. Cordon, J. *et al.* Identification of clinically relevant brain endothelial cell biomarkers in plasma. *Stroke* **54**, 2853 (2023).
13. Tin, A. *et al.* Identification of circulating proteins associated with general cognitive function among middle-aged and older adults. *Commun. Biol.* **6**, 1117 (2023).
14. Oh, H. *et al.* Organ aging signatures in the plasma proteome track health and disease. *Nat.* **624**, 164 (2023).
15. Dark, H. *et al.* Proteomic indicators of health predict alzheimer's disease biomarker levels and dementia risk. *Annals Neurol.* **95**, 260 (2024).
16. Candia, J. *et al.* Assessment of variability in the somascan assay. *Sci. Reports* **7**, 14248 (2017).
17. Kim, C. *et al.* Stability and reproducibility of proteomic profiles measured with an aptamer-based platform. *Sci. Reports* **8**, 8382 (2018).
18. Tin, A. *et al.* Reproducibility and variability of protein analytes measured using a multiplexed modified aptamer assay. *J. Appl. Lab. Medicine* **4**, 30–39 (2019).
19. Raffield, L. *et al.* Comparison of proteomic assessment methods in multiple cohort studies. *Proteomics* **20**, 1900278 (2020).
20. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun.* **12**, 6822 (2021).
21. Dubin, R. *et al.* Analytical and biological variability of a commercial modified aptamer assay in plasma samples of patients with chronic kidney disease. *Appl Lab Med* **8**, 491 (2023).
22. Candia, J., Daya, G., Tanaka, T., Ferrucci, L. & Walker, K. Assessment of variability in the plasma 7k somascan proteomics assay. *Sci. Reports* **12**, 17147 (2022).
23. Assarsson, E. *et al.* Homogenous 96-plex pea immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS ONE* **9**, e95192 (2014).
24. Cheung, F. *et al.* Web tool for navigating and plotting somalogic adat files. *J. Open Res. Softw.* **5**, 20 (2017).
25. Cotton, R. & Graumann, J. readat: An r package for reading and working with somalogic adat files. *BMC Bioinforma.* **17**, 201 (2016).
26. Johnson, W., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostat.* **8**, 118 (2007).
27. Leek, J., Johnson, W., Parker, H., Jaffe, A. & Storey, J. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma.* **28**, 882 (2012).

28. Reese, S. *et al.* A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinforma.* **29**, 2877 (2013).

29. Hong, M.-G., Lee, W., Nilsson, P., Pawitan, Y. & Schwenk, J. Multidimensional normalization to minimize plate effects of suspension bead array data. *J. Proteome Res.* **15**, 3473–3480 (2016).

30. Lim, S. *et al.* Evaluation of two high-throughput proteomic technologies for plasma biomarker discovery in immunotherapy-treated melanoma patients. *Biomark. Res.* **5**, 32 (2017).

31. Graumann, J. *et al.* Multi-platform affinity proteomics identify proteins linked to metastasis and immune suppression in ovarian cancer plasma. *Front. Oncol.* **9**, 1150 (2019).

32. Kukova, L. *et al.* Comparison of urine and plasma biomarker concentrations measured by aptamer-based versus immunoassay methods in cardiac surgery patients. *J. Appl. Lab. Medicine* **4**, 331–342 (2019).

33. Lopez-Silva, C. *et al.* Comparison of aptamer-based and antibody-based assays for protein quantification in chronic kidney disease. *CJASN* **17**, 350–360 (2022).