STEpUP OA SomaScan QUALITY ASSESSMENT PLAN

v2.0, 27/05/2022

This document describes the quality assessment measures that the Data Analysis Group will apply to all tranches of SomaScan data. The aim of this analysis is to produce a series of quality control measures to assess different aspects of the SomaScan's performance. These measures will be used to decide on a quality control procedure (including protein and sample exclusion criteria) for the full analysis plan going forward.

Definitions:
Plate: the 96 well plates on which SomaLogic processes samples (the majority of samples are randomly assigned to plates).
Tranche: a set of samples which are sent to SomaLogic for processing at the same time.
Plasma QC purposes.
Pool sample: A pooled synovial fluid sample that is added by STEpUP OA to all plates for QC purposes.
QC sample: A plasma sample of known concentration that is added by SomaLogic to all plates for QC purposes
CV: Coefficient of variation, i.e. the standard deviation of repeated samples divided by the mean across repeats.
Variance explained: Percent of variation in a particular sample type that is predicted to be driven by technical variation, measured by the ratio of variance in pooled samples to variance in all samples.

Total protein check
The simplest check will be the total protein signal for each sample, defined as the sum of all relative concentrations across all proteins for that sample. Outliers on this measure (with all proteins at abnormally high or low concentration) may represent poor quality samples.
Output: A histogram of total protein signal across samples. This will be carried out unstratified and stratified by injury status and blood staining grade.

Limit of detection check
For each protein on each plate we calculate a limit of detection (LoD), which is defined as the median concentration of the blank plus 4.9 times the median absolute deviation of the blank. For each sample, we calculate the proportion of proteins for which that sample is below LoD, and for each protein we calculate the proportion of samples that have concentrations below the LoD. We will also estimate a "rule-of-thumb" upper LoD (uLoD), defined as 80,000 RFU in the raw, unnormalized data.
Output: A histogram of proportions of proteins LoD for each sample and a histogram of proportions of samples LoD for each protein. Similar histograms for uLoD.

Checks against synovial fluid pool samples
Each plate includes two pool samples, each taken from one of two large pools of samples (an injury pool and an OA pool), which will be used to assess the quality of each plate and to measure plate-to-plate variation. In addition, one plate will include a second pair of pool samples to test intra-plate repeatability/reproducibility (within plate pool, inj and OA are on different plates), and another will include a pair of pool samples that have been repeatedly frozen and thawed 5 times to understand protein stability. Some plates will also include an

unspun pool (i.e. a pool of unspun synovial fluid samples), as well as multiple fresh-processed pools (i.e. pools that have been hyaluronidase treated at different times). Each plate will also include three plasma QC samples.
Outputs: Plots showing cumulative distributions of coefficient of variation (CV) and variance explained across all proteins using the pool samples.
We will also calculate mean CVs and mean variance explained for:

- interplate pools (i.e. pooled samples on different plates) - run separately for injury pool, spun OA pool and unspun OA pool.
- intraplate pools (i.e. pooled samples on the same plate).
- individual plates (i.e. the difference between the value for that plate vs the mean of values across all other plates).
- high-freeze-thaw and low-freeze-thaw samples
- pools treated at different times
- plasma QC samples

All measures will be broken down by injury vs OA vs unspun pool.

Principal component analysis (PCA)

We will use PCA and UMAP to visualize the main axes of correlated protein variation across samples, and to check visually for obvious signs of outlying samples, batch variation or other latent structures. These will later be checked against technical confounders (see below).
Outputs: A cumulative distribution of eigenvalues (expressed as variance explained per PC), and plots of top PCs (at least five) and UMAP coordinates (2-dimensional) colour coded by plate and other relevant technical covariates from Table 1 (with histograms of top PCs by variable for factor variables).

Checks against possible technical confounders

We will test for the effect of a number of possible technical confounding variables (Table 1) to see if they impact protein signatures (either individual measured protein concentrations or as proteome-wide parameters). We will test each confounder individually in a linear regression against the relative concentration of each assayed protein, as well as against the total protein signal and against the top 10 principal components. For variables that we expect to differ between cohorts (marked with “C” in Table 1) we will include cohort as a covariate.
Outputs: A table of regression outputs (effect size, p-value) of each confounder against the top 10 PCs and total protein concentration, and then a QQ plot for each of the confounders summarising the distribution of regression p-values across all protein concentrations.

● Plate (factor)
● Cohort name (factor)
● QC group (factor)
● Number of freeze-thaw cycles prior to receipt (factor: 0, 1-5, >5) C
● Specific number of freeze-thaw cycles (linear) C
● Day of processing (factor) C
● Processing batch (factor) C
● Age of sample (linear) C
● Volume of sample (linear) C
● Blood staining grade (factor) C
● Plate position (factor)
● Date of run (factor)
● Tranche (factor)

Table 1: List of technical variables to include in regression (variable data type in brackets). "C" indicates that analyses will be carried out conditional on cohort.

Checks of selected biomarkers against immunoassays

We will validate the SomaLogic quantification against protein concentrations measured in the same samples using Meso Scale Discovery (MSD) or other sandwich immunoassays which have been validated by the OA Centre translational lab for use with synovial fluid. These assays have been measured on up to 59 of the first tranche of samples, primarily on non-hyaluronidase treated samples, with two assays run on a small number of HAse-treated samples. 11 of these biomarkers are included on the SomaScan v4.1 platform.

Outputs: Pearson and Spearman correlation coefficients between SomaLogic relative concentration and absolute concentration from the immunoassays for each of the biomarkers, summarized as a table (with estimate and 95% CIs). Plot calibration, i.e. predicted R2 (based on technical variance explained using pools) vs actual R2 (based on immunoassay correlation).

Biomarker	**Type of assay**	**SomaScan V1.3k ID**	**Number of samples**	**Pre/post HAse treatment**
TIMP-1	MSD Ultra-Sensitive	SL000591	59	Pre
MMP-3	MSD Ultra-Sensitive	SL000524	59	Pre
IL-6	MSD V-PLEX	SL000087	59	Pre
IL-8	MSD V-PLEX	SL000039	59	Pre
MCP-1	MSD V-PLEX	SL000038	59	Pre
FGF2 (bFGF)	MSD V-PLEX	SL000004	59	Pre
Activin A	R&D Quantikine ELISA	SL001938	59	Pre
LTBP2	ABBEXA ELISA	N/A	46	Pre
TGFβ (TGFb1)	R&D Quantikine ELISA	SL000584	58	Pre
TSG-6	MSD SELF-COATED	SL004782	58	Pre
CD14	R&D Quantikine ELISA	SL018823	9	Post
VEGF	MSD V-PLEX	SL000002	9	Post

Table 2: Biomarkers to be used for validation

Comparison with older v4.0 v4.1 data

We have older data, of variable quality, on SomScan v4.0 and v4.1, on the same samples. The old tranche 1 data is v4.0, and has been assessed to be high quality. The old tranche 2 data is on v4.0, but is of low quality, and samples were not randomised across plates. The old tranche 3 data is of intermediate quality and is on v4.1, but again samples were not randomised across plates. We will assess the correlation between the old data and the new data, broken down by tranche. We will restrict our analysis to proteins present on v4.0 and

v4.1, with an additional analysis for tranche 3 of the additional 2k proteins present only on v4.1. We will also assess predictors of sample repeatability (injury/OA, cohort) within tranche.
Output: Histograms of correlation coefficients for each protein, by tranche. Boxplots of correlation coefficients for injury and OA samples separately, and for each cohort separately, broken down by tranche.

Temperature control experiment

Three OA synovial fluid samples were additionally processed at three different temperatures (low, normal and high[1]) to assess which proteins are sensitive to processing temperature, and to test whether processing temperature could be a cause of the dropout signal described in the next section. We will test which proteins differ significantly between the different processing temperatures using paired t-tests between the high and low conditions, and will also compare dropout signal markers for the three experimental conditions and for previous runs of the same samples.
Output: P-values for temperature sensitivity for each protein, and plots of values of dropout marker proteins across the experimental conditions and previous runs of the same samples.

Testing for bimodal protein signal

We know from past experiments that processing batch (i.e. batch in which samples were defrosted and enzyme treated) has a strong impact on protein concentration. This introduces a bimodal pattern of protein expression that varies across batches but tends to be consistent within batches (where a subset of proteins all have systematically higher or lower concentration within each processing batch). This bimodal signal is often a significant source of variation in protein concentration, and often drives its own principal component. We will select the top PCs associated with processing batch (conditional on cohort), and use these to define a binary division in the data. We will also test each protein for correlation with this bimodal signal in a linear model (conditional on cohort), in order to define a set of “bimodal-associated” proteins. We will also test whether individual proteins show patterns of bimodality using the Hartigan’s Dip Test.
Output: PCA and UMAP plots coloured by bimodal signal status. A QQ plot of p-values for bimodal-associated proteins. A QQ plot for p-values for the Hartigan’s Dip Test.

Spun vs unspun samples

In order to aid comparison of samples processed in different ways, a total of 18 samples were processed in pairs, with the first of the pair being centrifuged between aspiration and freezing (spun) and the second being frozen without centrifugation (unspun). We will use these samples to define proteins which are sensitive to centrifugation (with centrifugation-dependent proteins mostly likely representing cellular material). We will measure two readings of sensitivity: correlation (i.e. whether concentrations are correlated between spun and unspun samples) and differential abundance (i.e. whether spun and unspun samples have systematically different concentrations).
Output: A QQ plot of p-values for correlation (measured in a Pearson correlation test) and differential abundance (measured using a paired t-test). A plot of point estimates of correlation (pearson correlation coefficient) and normalized differential abundance (measured by Cohen’s d) for each protein.

Checking blood contamination data

[1] Described in the SOP “Re-run_Temperature\ processing\ controls_Vs\ 4_\ 24\ Dec\ 2021.pdf”

Blood contamination is likely to be a source of confounding in the proteomic data, particularly for the injury samples. Four biomarkers of blood contamination that are present on the SomaScan are hemoglobin (HBA1.HBB.4915.64), catalase, peroxiredoxin and carbonic anhydrase 1. We have also recorded a categorical blood staining grade, based on visual inspection prior to centrifugation, for a subset of samples. This analysis will test the correlation between visual- and biomarker-based measures of blood contamination.
Outputs: Scatterplots and correlation coefficients between protein blood contamination biomarkers and recorded blood staining grade, for all samples and broken down by injury vs non-injury cases.

Testing different normalisation approaches
We will test the impact of various different normalisation approaches on data quality and on principal components. These approaches will be:

- Unnormalized data
- The normalization approach validated in tranche 1 (hybNorm+plateScale+Calibration)
- SomaLogic's ANML approach (i.e. the fully normalised data provided by SomaLogic in the newest release)
- Concentrations batch corrected using ComBat by plate
- Concentrations batch corrected using ComBat by processing batch
- Concentrations batch corrected using ComBat by bimodal signal status

Output: Mean CV and predicted R2, QQ plot of p-values for bimodal-associated proteins, PCA coloured by confounders in Table 1, for each normalisation approach.

Checking sex-linked biomarkers
A number of protein biomarkers in plasma are known to differentiate between males and females, though for many of these proteins (e.g. prostate-specific antigen) these differences become more pronounced with age. It is unknown the extent to which these biomarkers distinguish sex in synovial fluid, and whether they can be used to test for sex mismatches. This analysis will validate these biomarkers in SF, and test to see if they can provide accurate prediction of reported sex.
We will test for known (serum) protein biomarkers of sex, including PSA (higher in males) and FSH, LH and beta HCG (all higher in females). We will validate that these show the expected relationship with recorded sex. To test predictive accuracy, we will fit a logistic regression classifier to the relative concentration data and produce class membership probabilities for each sample using cross-validation. Models will be fitted separately in under 50 and over 50 cohorts, and will be fitted separately in OA and injury samples. We will test whether certain individual plates have higher than expected sex mismatch probabilities, which could be indicative of a plating error.
Outputs: Violin plots of relative concentration of PSA, FSH, LH and beta HCG by recorded sex, both across all samples and broken down by age. Histograms of probabilities of sex mismatch for recorded males and females, broken down by plate and tranche.

Selection of outlier samples and proteins
An important output of this QC analysis will be to decide whether and how to modify our exclusion thresholds established in the tranche 1 and tranche 2 QA procedures. To assess our options for exclusions, we will explore per-protein filters on CV and variance explained (from pooled samples), p-value of association with technical confounders and on p-value for the bimodal-associated protein test. We will also explore per-sample filters on total protein and principal components.

Outputs: The number of proteins and samples removed for different thresholds of each filter. Lists of proteins and samples removed under illustrative examples of thresholds. Visualise impact of filters using UMAP and PCA.

Appendix 1: Variables for quality assessment

Variable	Source of data
● Plate (factor)	SomaLogic metadata
● Cohort name (factor)	STEpUP OA RedCap database
● QC group (factor)	STEpUP OA RedCap database
● Number of freeze-thaw cycles prior to receipt (factor: 0, 1-5, >5) C	STEpUP OA RedCap database
● Specific number of freeze-thaw cycles (linear) C	STEpUP OA RedCap database
● Day of processing (factor) C	STEpUP OA RedCap database
● Processing batch (factor) C	STEpUP OA RedCap database
● Age of sample (linear) C	STEpUP OA RedCap database
● Volume of sample (linear) C	STEpUP OA RedCap database
● Blood staining grade (factor) C	STEpUP OA RedCap database
● Plate position (factor)	SomaLogic metadata
● Date of run (factor)	SomaLogic metadata
● Tranche (factor)	STEpUP OA RedCap database