somalogic™

SomaScan® v4 Data Standardization and File Specification Technical Note

Definition of processes used to remove assay and sample bias from SomaScan v4 data and file specification for SomaScan results

1 Overview

Normalization and calibration are routine numerical procedures developed to remove systematic biases in the raw assay data. Normalization is a sample-by-sample adjustment in overall signal within a single plate (run) performed across three non-consecutive steps: Hybridization Control Normalization, Intraplate Median Signal Normalization, and Median Signal Normalization to a Reference. Plate Scaling and Calibration is a SOMAmer® binding reagent-by-SOMAmer binding reagent adjustment that minimizes between-plate variability. Global reference standards are established for procedures with controls on each plate. Individual, QC, and Calibrator samples are normalized and calibrated to the established global reference standards. Separate calibrator global reference standards are established for each matrix (serum, plasma), and assay shifts or skew from the global reference standards are tracked over time. New global reference standards may be developed in concordance with changes in assay processes, performance, or reagents.

Hybridization Control Normalization was developed to remove systematic biases present in the raw data after slide feature aggregation from a slide-based hybridization microarray for assay readout and quantification. Hybridization Control Normalization is performed using a set of twelve hybridization control sequences measured independently for each sample array. The procedure is intended to correct for systematic effects on the data introduced during the hybridization readout and results in a single scale factor for each sample that is subsequently applied to the measured signal on all features within a subarray (sample).

Intraplate Median Signal Normalization uses all the SOMAmer reagent signals on a given subarray to remove sample or assay biases that may be due to differences between samples in overall protein concentration, pipetting variation, variation in reagent concentrations, assay timing, and any other source of systematic variability within a single plate. Each SOMAmer reagent is assigned to one of three dilution sets, scale factors are derived within dilution sets separately, and all SOMAmer binding reagents within each set are scaled together. Three sample dilutions will result in three independent median signal scale factors for each subarray (sample) in addition to the hybridization scale factor. This step is only applied to calibrator samples.

Plate Scaling and Calibration is accomplished using a number of replicate measurements of a common pooled calibrator sample consistent with the assay sample type for a study. Calibrator samples must be composed of identical sample matrices as the samples that are being calibrated. No protein spikes are added to the calibrator samples – SomaLogic relies solely on the endogenous levels of each analyte within a calibrator sample. Since calibration attempts to correct plate-to-plate variation and such variation can be idiosyncratic for SOMAmer binding reagents, a unique calibration scale factor is derived for each SOMAmer binding reagent within the assay. The median of these scale factors is then computed and applied across all SOMAmer measurements in that plate to account for the total signal difference (plate scale), and the scale factors are subsequently recalculated for each SOMAmer and applied to all measurements within the set of samples in that plate.

Median Signal Normalization to a Reference occurs on a per-sample basis, wherein a scale factor for a set of SOMAmer reagents is computed against a reference value generated from a cohort of healthy normal individuals and then aggregated within a dilution. The median of each dilution's scale factors is then applied to their respective SOMAmer reagents. This step is applied to QC, Buffer, and individual samples.

File specification: SomaScan results are produced in a tab-delimited ASCII file with an ADAT extension (ADAT file) as described in section 6 below.

2 Normalization

2.1 Hybridization Control Normalization

A set of hybridization control sequences is added to each sample as part of the elution buffer in the SomaScan assay. These hybridization controls are added at concentrations to give measured relative fluorescence units (RFU) that span the dynamic range of the assay. The global reference RFU value for each hybridization control is defined by the median signal measured within the current plate being normalized.

A ratio is computed by dividing the median RFU for each control by its measured RFU in the sample. The median of these hybridization control measurement ratios in each subarray defines the sample-based hybridization scale factor. By definition, such a scaling will equate the median RFU for the hybridization controls to the median reference RFU for the controls. All SOMAmer reagent results within a sample are multiplied by the resulting hybridization scale factor increasing or decreasing the overall "brightness" of the sample. The procedure is displayed graphically in Figure 1.

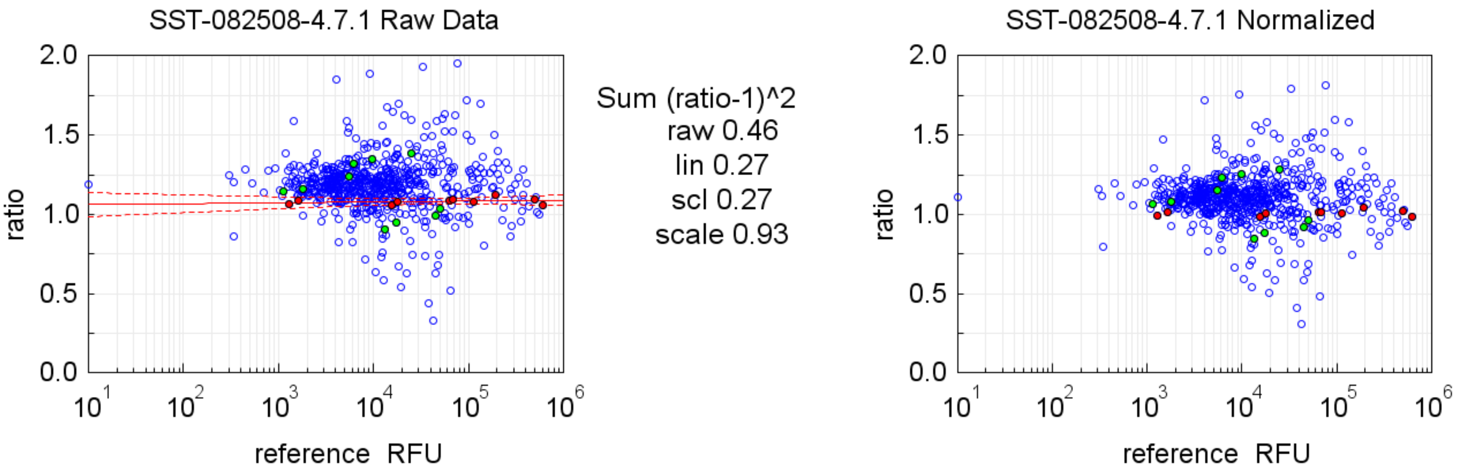

FIGURE 1 Scatter plot for sample ratios versus their reference RFU for Hybridization Control Normalization. The red filled circles are the RFU ratios for the hybridization controls fit with a regression line (red) displayed on the left plot prior to normalization. The inverse of the median ratio of the hybridization controls, 1/1.07, defines the scale factor = 0.935. The normalized data to which the 0.93 scale factor has been applied is displayed on the right plot. All ratios are decreased by application of the common scale factor and the controls now have a median ratio of one.

2.2 Intraplate Median Signal Normalization

Intraplate Median Signal Normalization is performed on each sample dilution independently. In most matrices, each SOMAmer binding reagent is assigned to one of three dilution sets, scale factors are derived within dilution sets separately, and all SOMAmer reagents within each set are scaled together. Within each sample matrix, this is only performed on calibrator samples. Like Hybridization Control Normalization which uses a local reference standard, the local median reference RFU for each SOMAmer reagent is the median RFU for that SOMAmer binding reagent within the sample group (calibrator in buffer) in the plate to be normalized. As in hybridization normalization, a ratio is computed for each SOMAmer reagent by dividing the reference SOMAmer RFU by its measured RFU in the sample to be normalized. The median of the SOMAmer measurement ratios for all SOMAmer reagents in a dilution defines the sample-based scale factor for all SOMAmer reagents within that dilution and sample. All SOMAmer reagents within the dilution for a sample are scaled by the resulting median signal scale factor. Three sample dilutions will result in three independent median signal scale factors for each sample in addition to the hybridization scale factor as shown in Figure 2.

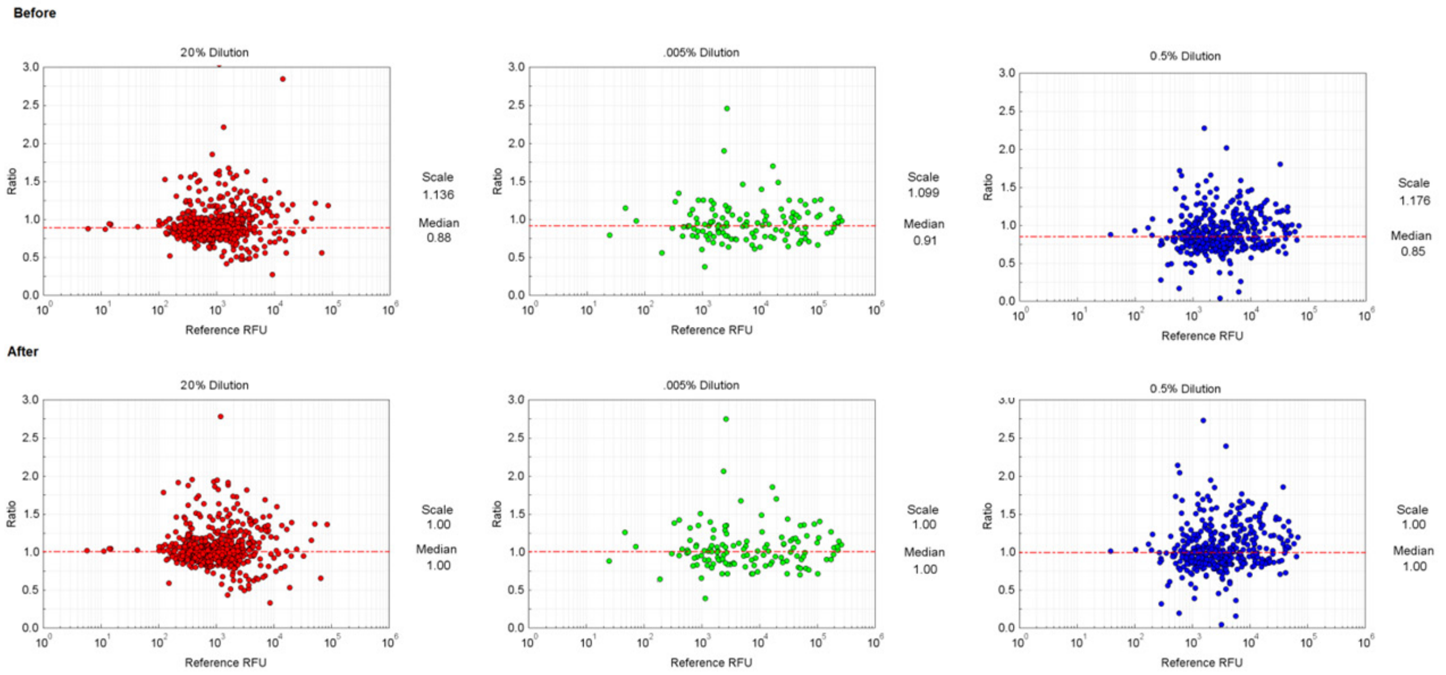

FIGURE 2 Scatter plots for sample ratios versus their reference RFU for median signal normalization. The filled circles are the RFU ratios for the SOMAmer binding reagents within a single dilution mix. The inverse of median ratios (1/0.88, 1/0.91 1/0.85) define the scale factors (1.136, 1.099, 1.176) applied to all the SOMAmer reagent measurements in the dilution mix, 20%, 0.005%, 0.5% respectively, for this sample. The ratios for the normalized data are displayed on the right. After normalization all ratios within the dilution mix are changed by application of the common scale factor and the SOMAmer reagents now have a median ratio of one, although the overall profile of the ratios defined by their relative distances has not changed.

3 Plate Scaling & Calibration

Clinical sample studies are plate scaled and calibrated to remove systematic assay variability. A set of control calibrator samples is used to detect and remove systematic variability between independent assay plates. Calibrator samples must be of the same type as the samples that are being calibrated. Calibrator global reference RFU values for each SOMAmer reagent are defined by the median signal measured on a set of samples spanning a number of independent assay plates that have been shown to meet assay acceptance criteria. For each SOMAmer reagent, the median RFU signal for that SOMAmer reagent across all the calibrator samples within the clinical study defines the global calibrator reference for that SOMAmer binding reagent. Figure 3 below displays the data from a typical clinical study and illustrates the systematic bias removed by calibration.

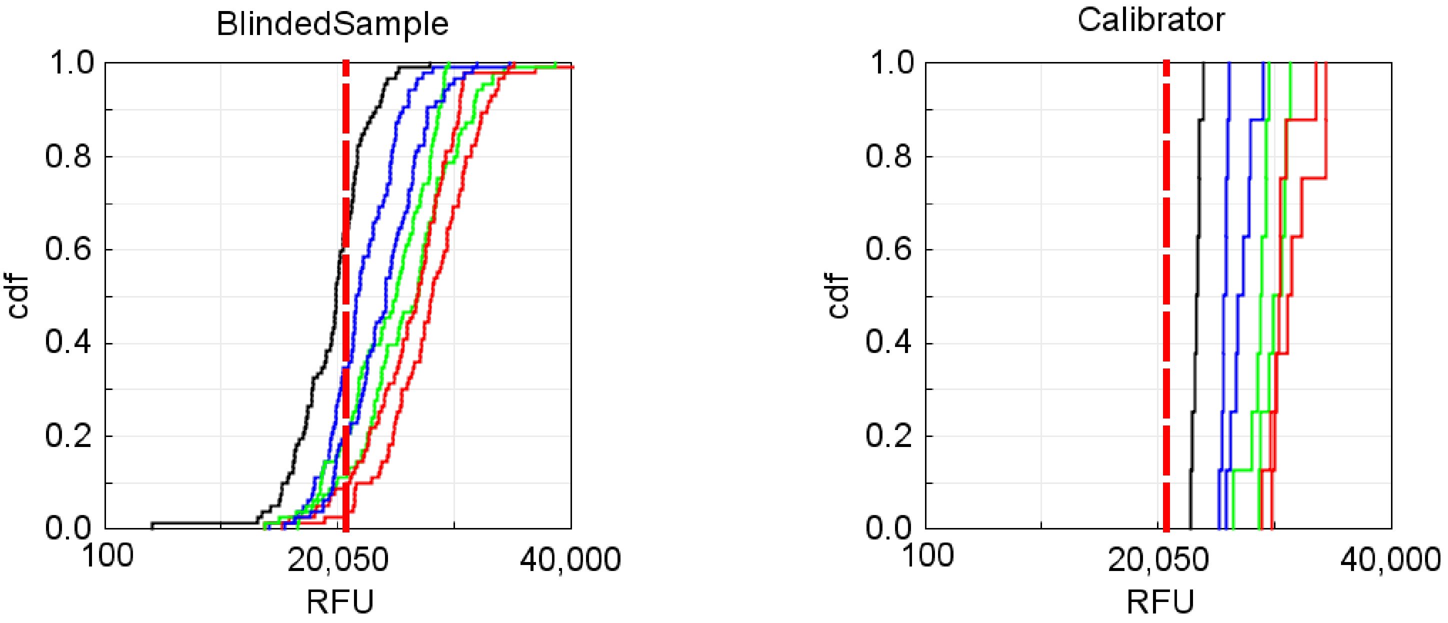

FIGURE 3 Sample distributions illustrating systematic bias between assay plates. Cumulative distribution functions (cdf) were generated for a single SOMAmer measurement across seven independent assay plates (approximately 600 samples randomly distributed across the plates) and are color coded by plate (left plot). The cdfs for the replicate calibrator sample measurements for each plate are displayed on the right for that SOMAmer reagent, color coded as on the left. The vertical red bar is the calibrator global reference obtained from a separate independent set of calibrator plates and is the target calibration RFU for this SOMAmer reagent's measurements. Note correlation of shifts between the clinical sample cdfs and the calibrator sample cdfs.

Plate scaling is performed on an entire independent plate. A local median reference value is derived for each SOMAmer reagent by computing the median RFU for that SOMAmer reagent from the set of replicate calibrator samples within the plate. The SOMAmer-based calibration scale factor is then computed by dividing the calibrator global reference RFU by the local median reference value defined for each SOMAmer reagent. The median of all scale factors for a given plate is then applied across all SOMAmer measures in the plate, forcing the overall calibrator median signal to match the overall median signal within the global calibrator reference.

Plate-to-plate calibration is performed on each SOMAmer measurement within the plate independently. A local median reference value is derived for each SOMAmer reagent by computing the median RFU for that SOMAmer reagent from the set of replicate calibrator samples within the plate. The SOMAmer-based calibration scale factor is then computed by dividing the calibrator global reference RFU by the local median reference value defined for each SOMAmer reagent. This scale factor is applied to all SOMAmer measurements in the plate, forcing the median calibrator signal to match the global calibrator reference for that SOMAmer binding reagent. Each plate within a study has a unique calibration scale factor for each SOMAmer reagent. The data from Figure 3 are displayed after calibration in Figure 4 below.

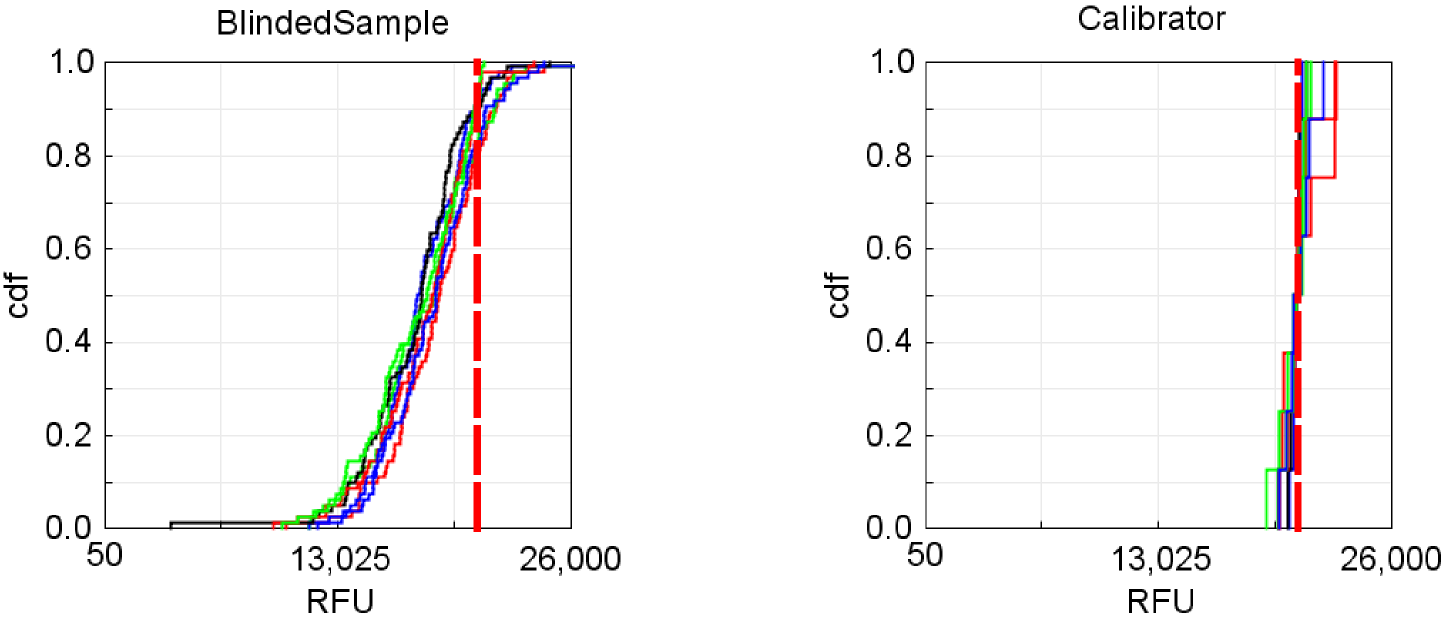

FIGURE 4 Sample distributions illustrating removal of systematic bias between assay plates. Cumulative distribution functions (cdf) for a set of clinical samples from Figure 3 after calibration (left plot). The cdfs for the replicate calibrator sample measurements for that SOMAmer reagent are displayed on the right. The median of each calibrator sample distribution equals the calibrator global reference standard by definition. Note the collapse of the clinical sample distributions to essentially a single distribution after calibration.

4 Median Normalization to a Reference

All individual, QC, and Buffer samples are then median normalized to a reference value. Unlike Intraplate Median Signal Normalization, Median Normalization to a Reference can be performed on a single sample due to the presence of an external global reference value generated from a cohort of healthy normal individuals for each SOMAmer reagent for our core matrices. This method is very similar to Intraplate Median Signal Normalization in practice, the primary difference being the origination of the reference value. A ratio is computed for each SOMAmer reagent by dividing the global reference SOMAmer RFU by its measured RFU in the sample to be normalized. The median of the SOMAmer measurement ratios for all SOMAmer reagents in a dilution defines the sample-based scale factor for all SOMAmer reagents within that dilution and sample. All SOMAmer reagents within the dilution for a sample are scaled by the resulting median signal scale factor. Three sample dilutions will result in three independent median signal scale factors for each sample in addition to the hybridization scale factor. We then iterate this approach up to 100 times until convergence occurs. Only ratios within 2 standard deviations of the mean will be considered for calculating scale factors. This approach is known as Adaptive Normalization by Maximum Likelihood, or ANML. If the samples are a non-core matrix, we assemble an intra-study reference via bootstrapping and calculate the ratios from that reference to individual samples.

5 Acceptance Criteria

Hybridization Control and Intraplate Median Signal Normalization scale factors are expected to be in the range of 0.4-2.5. The plate scale factor is expected to be between 0.4 and 2.5. The distribution of QC sample ratios is expected to have 85% of individual SOMAmer reagents in the total array between 0.84 and 1.19 (i.e. less than 15% in the tails of the distribution). Gaussian distributions of scale factors are expected. A report is provided for each study (single plate or set of plates) with the results of the Normalization and Calibration process.

6 File Specification

SomaScan results are produced in a tab-delimited ASCII file with the filename extension ".adat" – an ADAT file. The ADAT file contains measurements for a series of analytes (columns) across a series of samples (rows) and includes analyte description and sample description information. The format is designed to provide flexibility for the number of samples as well as the number and types of analyte and sample descriptors.

6.1 File Type

Deliverables will be created as tab-delimited ASCII files and include the sections described in 6.2.

6.2 File Content

6.2.1 ADAT File Structure and Layout

The ADAT file is divided into 4 sections: Header Information, Column Data Description, Row Data Description, and Data Table Section. The Data Table Section is composed of three subsections: Analyte Description Information/Column Metadata, Sample Description Information/Row Metadata, and the SomaScan Assay Results section. Figure 5 shows an ADAT file opened in a text editor with each section highlighted and labelled.

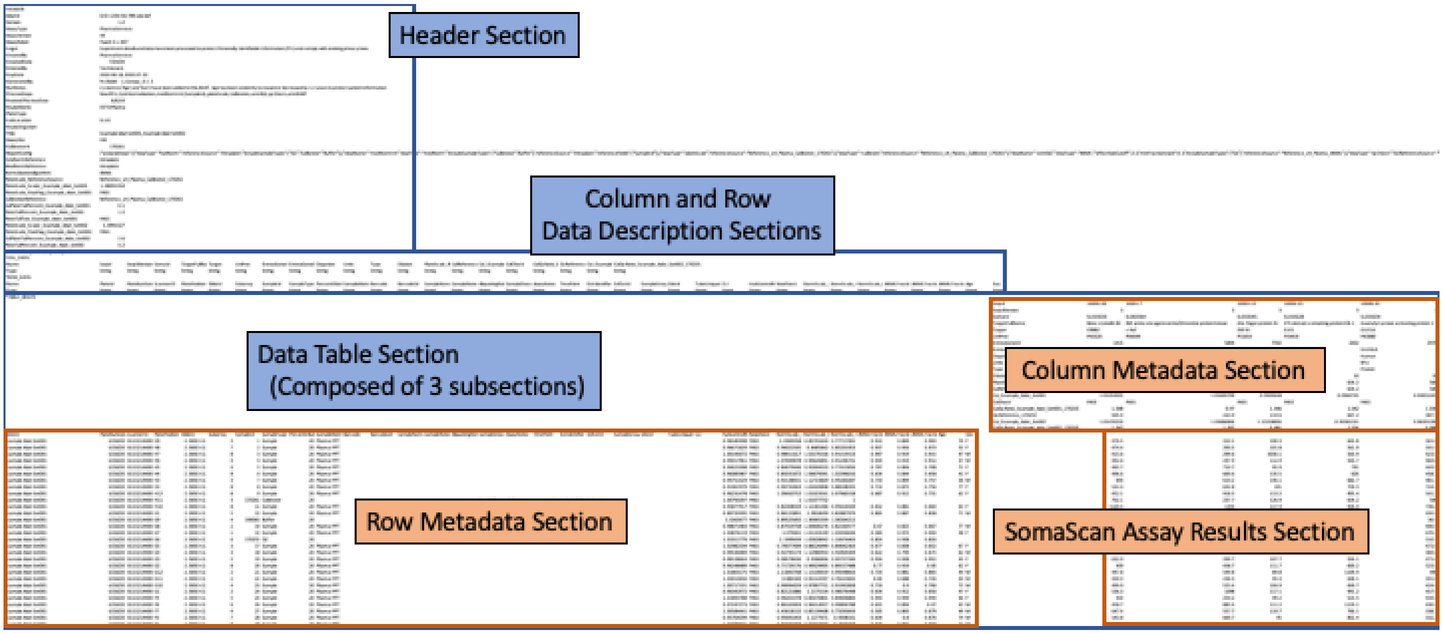
Header Section
Column and Row Data Description Sections
Data Table Section (Composed of 3 subsections)
Column Metadata Section
Row Metadata Section
SomaScan Assay Results Section

FIGURE 5 An ADAT file is shown opened in a text editor and each section labelled to illustrate the organization and layout of the file.

6.2.2 Header Information

This section starts following the row containing the entry “^HEADER” and consists of two columns separated by a tab. The Header Section provides plate and file metadata including references used, process steps, experiment notes as recorded by the assay services team, experiment identifiers used in creation of the file, summary scale, QC metrics by plate, and other information related to the particular SomaScan assay.

6.2.3 Column and Row Data Descriptions

The Column and Row Data Descriptions follow rows containing the entries “^COL_DATA” and “^ROW_DATA”, respectively. Each section consists of two rows. The first column contains row headers “!Name” and “!Type”.

- The “!Name” row of the “^COL_DATA” section contains the expected names of the columns present in the Analyte Description Information/ Column Metadata section (see section 6.2.4.1), while for the “^ROW_DATA” section it contains the expected names of the columns present in the Sample Description Information/Row Metadata section (see section 6.2.4.2).
- The “!Type” row in both cases was historically used to indicate expected data types present in each data column, but typically contains the entry “String” for each column currently.

6.2.4 SomaScan Data Table

The SomaScan Data Table begins following the row containing the entry "^BEGIN_TABLE" and is composed of three subsections: Analyte Description Information/Column Metadata, Sample Description Information/Row Metadata, and the SomaScan Assay Results section.

6.2.4.1 Analyte Description Information/ Column Metadata

This section provides multiple descriptors for each SOMAmer Reagent and begins in the first row following the "^BEGIN_TABLE" entry. The starting column will depend on the number of columns present in the Sample Description Information/Row Metadata section, and thus will be preceded by a number of empty entries in the file. The first Analyte Description column contains field names for each descriptor and each subsequent column contains descriptors associated with the SOMAmer Reagent associated with that column. In the event that more than one descriptor is required for a given field (e.g. EntrezGeneID, UniProt), a space or semicolon will be used as a delimiter for successive entries.

Field	Description	Example
SeqId	SomaLogic sequence identifier	2182-54_1
SeqidVersion	Version of SOMAmer sequence	2
SomaId	Target identifier, of the form SLnnnnnn (8 characters in length)	SL000318
TargetFullName	Target name curated for consistency with UniProt name	Complement C4b
Target	SomaLogic Target Name	C4b
UniProt	UniProt identifier(s)	P0C0L4 P0C0L5
EntrezGeneID	Entrez Gene Identifier(s)	720 721
EntrezGeneSymbol	Entrez Gene Symbol names	C4A C4B
Organism	Protein Source Organism	Human
Units	Relative Fluorescence Units	RFU
Type	SOMAmer target type	Protein
Dilution	Dilution mix assignment	0.01%
PlateScale_Reference	PlateScale reference value	1378.85
CalReference	Calibration sample reference value	1378.85
medNormRef_ReferenceRFU	Median normalization reference value	490.342
Cal_V4-<YY>-<SSS>-<PPP>	Calibration scale factor (for given year-study-plate)	0.64
ColCheck	QC acceptance criteria across all plates/sets	PASS
QcReference_<LLLLL>	QC sample reference value (for given QC lot)	PASS
CalQcRatio_V4-<YY>-<SSS>-<PPP>	Post calibration median QC ratio to reference (for given year-study-plate)	1.04

6.2.4.2 Sample Description Information/ Row Metadata

This section provides multiple descriptors for each sample and begins in the first column of the file. The starting row will depend on the number of rows present in the Analyte Description Information/Column Metadata section, and is the first row after the last analyte description field. The first Sample Description row contains column names for each descriptor and each subsequent row contains descriptors associated with the sample.

Field	Description	Example
PlateId	Plate identifier	V4-18-004_001,V4-18-004_002
ScannerID	Scanner used to analyze slide	SG12064173,SG14374437
PlatePosition	Location on 96 well plate (A1-H12)	A1, H12
SlideId	Agilent slide barcode	258495812746
Subarray	Agilent subarray (1 – 8)	1,8
SampleId	1st form is Subject Identifier, 2nd form (calibrators, buffers)	2031
SampleType	1st form for clinical samples (Sample), 2nd form as above	Sample, QC, Calibrator, Buffer
PercentDilution	Highest concentration the SOMAmer dilution groups	20
SampleMatrix	Sample matrix	Plasma-PPT
Barcode	1D Barcode of aliquot	S622225
Barcode2d	2D Barcode of aliquot	191055125
SampleName	Internal sample identifier	S0000-5555555, Blank, Calibrator, QC
SampleNotes	Assay team sample observation	Cloudy, Low sample volume, Reddish
AliquotingNotes	Assay team aliquot observation	Short 2nd
SampleDescription	Supplemental sample information	Plasma QC 1
AssayNotes	Assay team run observation	Beads aspirated, Leak/Hole, Smear
TimePoint	Sample time point	Baseline
ExtIdentifier	Primary key for Subarray	EXID40000000032037
SsfExtId	Primary key for sample	EID102733
SampleGroup	Sample group	A, B
SiteId	Collection site	SomaLogic, Covance
TubeUniqueID	Unique tube identifier	112288990011
CLI	Cohort definition identifier	CLI6006F001
HybControlNormScale	Hybridization control scale factor	0.94830449
RowCheck	Normalization acceptance criteria for all row scale factors	PASS, FLAG
NormScale_0_5	Median signal normalization scale factor (0.5% mix)	1.02718047
NormScale_0_005	Median signal normalization scale factor (0.005% mix)	1.11975411
NormScale_20	Median signal normalization scale factor (20% mix)	0.99614766
ANMLFractionUsed_0_5	Percentage of measurements used for ANML scale factor calculation (0.5% mix)	0.965
ANMLFractionUsed _0_005	Percentage of measurements used for ANML scale factor calculation (0.005% mix)	0.941
ANMLFractionUsed _20	Percentage of measurements used for ANML scale factor calculation (20% mix)	0.937

6.2.4.3 SomaScan Assay Result Information

Assay results begin at the second row after the last Analyte Descriptor and the second column after the last Sample Descriptor. Results are delivered in relative fluorescence units (RFU) and all results are numeric and formatted to one decimal place. Note that there is a bank row following the Analyte Description Information/ Column Metadata section and a blank column following the Sample Description Information/ Row Metadata section. These are included in order to facilitate the row headers of the Column Metadata and the column headers of the Row Metadata, maintaining alignment of the samples and the analytes in the SomaScan Assay Results.