1. Executive Summary

Method development in response to the requirement of population-based reference normalizing single samples with unknown sample handling artifacts and unknown clinical conditions or interventions. This method was shown to reject common sample handling markers from normalization, showed good concordance in scale factors to population-based adaptive median normalization (PBAMN), and retained bio-marker effect sizes.

Adaptive Normalization by Maximum Likelihood (ANML) was developed to be used on all applicable SOMAscan assay data including QC samples. This is an adaptive procedure which censors analytes during normalization that fall beyond the expectations for a measurement from our reference distribution and calculated dilution specific scale factors on individual samples.

Under ANML, the scale factor is calculated by maximizing the probability that a sample's RFU measurements came from the reference distribution. This is accomplished by using the information of the reference population variance for each analyte in addition to the reference distribution median. It can be shown that median normalization is an approximation to ANML in the limit the variance of the population is unknown. The analytes that are chosen to be excluded during normalization are based on the information contained within individual samples; that is, the scale factor for each sample would not change if new samples were brought in during insight development.

Table of Contents

Contents

2. Method Specification

ANML is an adaptive procedure which censors analytes during normalization that fall beyond the expectations for a measurement from our reference distribution and calculated dilution specific scale factors on individual samples.

2.1. Derivation

The objective of ANML is to maximize the probability that a given sample came from a certain reference distribution. A scale factor, α, is calculated that is applied to all analyte measurements within a dilution to maximize this probability, assuming they are derived from a normal distribution.

For a single measurement, the probability of observing this measurement given the reference distribution with respect to a scale factor is described by:

$$P(\alpha) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\frac{(\alpha + x - \mu)^2}{\sigma^2}\right] \quad (1)$$

Where for SOMAscan data, α, x and μ and implied in $\log_{10}$ space. To optimize this probability:

$$\alpha + x - \mu = 0$$
$$\alpha = \mu - x$$

Writing the logarithm out explicitly, we can get the scale factor in RFU space parameters:

$$\log(\tilde{\alpha}) = \log(\tilde{\mu}) - \log(\tilde{x})$$
$$\tilde{\alpha} = \frac{\tilde{\mu}}{\tilde{x}}$$

For more than one analyte this probability is a product over the individual measurements:

$$P(\alpha) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left[-\frac{1}{2}\frac{(\alpha + x_i - \mu_i)^2}{\sigma_i^2}\right] \quad (2)$$

To solve for the appropriate scale factor, we take the natural logarithm of both sides, take the first derivative with respect to α and set the resulting equation equal to 0 to solve for α.

$$Ln(P(\alpha)) = \sum_{i=1}^{N} Ln\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) - \frac{1}{2}\frac{(\alpha + x_i - \mu_i)^2}{\sigma_i^2}$$

$$\frac{dLn\big(P(\alpha)\big)}{d\alpha} = \frac{d}{d\alpha}\left(\sum_{i=1}^{N} Ln\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) - \frac{1}{2}\frac{(\alpha + x_i - \mu_i)^2}{\sigma_i^2}\right) = 0$$

$$0 = \sum_{i=1}^{N} \frac{-\alpha - x_i + \mu_i}{\sigma_i^2} \frac{1}{2\alpha \ln(10)}$$

$$\sum_{i=1}^{N} \frac{\alpha}{\sigma_i^2} = \sum_{i}^{N} \frac{\mu_i - x_i}{\sigma_i^2}$$

$$\alpha = \frac{\sum_{i=1}^{N} \frac{\mu_i - x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

Explicitly writing the parameters as logarithms we get, in RFU space:

$$\log_{10}(\tilde{\alpha}) = \frac{\sum_{i}^{N} \frac{\log_{10}(\tilde{\mu}_i) - \log_{10}(\tilde{x}_i)}{\sigma_i^2}}{\sum_{i}^{N} \frac{1}{\sigma_i^2}} \quad (3)$$

The appropriate scale factor is then the exponentiation of equation 3 to the base of 10.

2.2. Recovery of Median Normalization

In the situation that the variance of an analyte measurements is unknown and all σ are set equal, equation 3 reduces to:

$$\log_{10}(\tilde{\alpha}) = \frac{\sum_{i}^{N} \log_{10}(\tilde{\mu}_i) - \log_{10}(\tilde{x}_i)}{N}$$

$$\log_{10}(\tilde{\alpha}) = \frac{1}{N} \sum_{i=1}^{N} \log_{10}\left(\frac{\tilde{\mu}_i}{\tilde{x}_i}\right) \quad (4)$$

The calculated scale factor from equation 4 would be the average of the reference to sample ratios across all analytes. Median normalization (where the median of the ratios is used, instead of the average) is then an approximation to ANML when the variance of all analytes is unknown.

2.3. Performance

The median of scale factors is closer to unity and the distribution of scale factors is less extreme than those generated using median normalization to the reference. In reduction of assay variance as estimated by QC

control replicates, ANML application performs in an equivalent fashion to median normalization to the reference. Compared to median normalization to a reference: bias in overall scaling is reduced for clinical samples, sensitivity to change in overall signal from biological or pre-analytical variance is reduced.

The count of samples flagged as outside of the acceptance range may be reduced but extreme assay failures as characterized by median normalization scaling outside of the range of 0.25 - 4.0 will be consistently identified using ANML metric acceptance criteria to be determined in verification and validation.

2.4. Reference Dataset

Covance, CLI6006F001, used for generation of point and variance estimates for the ANML reference.

Dataset Name	Inclusion Criteria	Exclusion Criteria
Covance (n=1029) Collected 2008	• Males or females, between 20 and 80^{+} years of age, inclusive • No history of problems with blood draws, and assessment that veins will allow successful blood draws • Being able to comprehend and willing to sign an Informed Consent Form (ICF).	• Uncontrolled hypertension (i.e., 2 measures > 160/95, 10 minutes apart) • Self-reported treatment for a malignancy other than squamous cell or basal cell carcinoma of the skin in the last 2 years • Self-reported pregnancy • Self-reported chronic infectious (e.g., hepatitis B, hepatitis C, HIV), autoimmune, or other inflammatory condition(s) such as SLE, scleroderma, MS, Crohn's Disease, or ulcerative colitis • Self-reported chronic kidney or liver disease, chronic heart failure or diagnosed with myocardial infarction in the last 3 months, self-reported uncontrolled diabetes (HbA1c > 8% if known) • Self-reported acute viral or bacterial infection or a temperature >38°C within 24 hours of enrollment • Self-reported participation in any therapeutic study in the 14 days prior to blood sampling • Taking more than 20 mg/day of prednisone or related drug (self-reported)

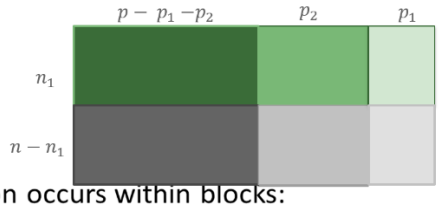
Covance		Site						
		Total	Austin	Boise	Dallas	Honolulu	Portland	San Diego
Total	*Count*	1029	84	216	52	168	252	257
Age	*Mean*	50.7	57.0	45.6	60.0	57.0	49.4	48.4
	Stdev	17.23	10.46	16.16	9.74	18.42	19.46	15.49
	Min	19	40	20	41	19	20	20
	Max	89	78	87	86	88	89	82
Sex	*F*	569	49	126	35	100	126	133
	M	460	35	90	17	68	126	124
TTS	*Mean*	3.6	3.2	0.9	5.2	4.3	4.6	4.3
TTD	*Mean*	3.0	2.3	1.4	4.8	4.7	3.2	3.0

3. Normalization methods

Normalization scaling for all non-ANML normalization steps may be generalized as the median of a vector of ratios: [reference/sample].

SOMAmer groups are most commonly defined by dilution bin (reported *ColMeta*: Dilution [0, 20, 0.5, 0.005] for SomaScan assay V4 plasma and serum; and Sample groups are most commonly defined by sample type (reported *RowMeta*: SampleType [Calibrator, QC, Buffer, Sample].

- Data Matrix (X_{nxp}):
  - Samples and SOMAmers have natural groupings



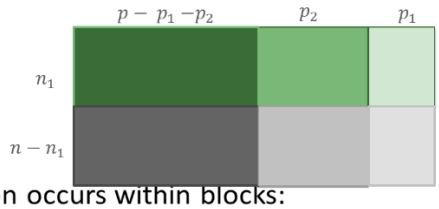
  - Normalization occurs within blocks:
    - SOMAmer groups: column index $j \in \mathbb{J}$
    - Sample groups: row index $i \in \mathbb{I}$

Commonly used reference values:

- Hybridization normalization: within plate, within dilution 0, for each Hybridization Control Elution SOMAmer, median signal from control (Calibrator, QC, Buffer) replicates on the plate. Used with V4 SomaScan Assay data (hybridization control signal from all samples and controls on a plate with SomaScan Assay data)
- Median signal normalization (controls): within plate, for each SOMAmer within non-zero dilution bin, within SampleType (Calibrator, Buffer), median signal from replicates on the plate. Used with V3 and V4 SomaScan Assay data.
- Median signal normalization (samples): within plate, for each SOMAmer within non-zero dilution bin, within SampleType (Sample), median signal from samples on the plate Used with V3 SomaScan Assay data (randomization requirement for V3 SomaScan Assay).
- Median signal normalization to a study reference: across plates, for each SOMAmer within non-zero dilution bin, within SampleType (Sample), median signal from samples in the study. Used alternatively with V3 SomaScan Assay data and with V4 non-core specimen types.
- Median signal normalization to a population reference: independent runs of assay plates for reference population, for each SOMAmer within non-zero dilution bin, within SampleType (Sample), median signal from samples in the study. Used in as early V4 standardization method.
- Adaptive Normalization using Maximum Likelihood (to a population reference): independent runs of assay plates for reference population, for each SOMAmer within non-zero dilution bin, within SampleType (Sample), median signal and robust variance of signal (MAD) from samples in the study. SomaScan V4 references only.

Normalization scaling can be generalized with the following formulation:

- For each $i \in \mathbb{I}, j \in \mathbb{J}$
  - $x_{i,j}$ = Signal on j^{th} SOMAmer in the i^{th} sample
  - r_j = Reference Level for j^{th} SOMAmer
  - j^{th} Reference ratio: $\rho_{i,j} = \left(\frac{r_j}{x_{i,j}}\right)$
- Scale Factor : $s_i = median_{j \in \mathbb{J}} \ (\rho_{i,j})$
- Normalized Data: $\acute{x}_{i,j} = s_i * x_{i,j}$
  - In matrix form: $\acute{X}_{nxp} = S_{nxn} X_{nxp}$
    - $S_{nxn} = diag(s), \quad X_{nxp} = [x_{i,j}] \ i \in \mathbb{I}, j \in \mathbb{J}$
- Median and Hyb Normalization schemes differ in the specification of
  - Sample groups ($\mathbb{I}$)
  - SOMAmer groups ($\mathbb{J}$)
  - Reference signal levels (r)
- Data Matrix (X_{nxp}):
  - Samples and SOMAmers have natural groupings



  - Normalization occurs within blocks:
    - SOMAmer groups: column index $j \in \mathbb{J}$
    - Sample groups: row index $i \in \mathbb{I}$