Cheat Sheet

Prefilter

Option	*Action*
x% permissible missing QCs	Metabolite feature is removed if there are more than x% missing QC samples.
☒ Batch-wise	% missing calculated per batch and remove peak if any batch has > x% missing QCs
☒ < blank = missing	consider any QC sample with signal < the blank threshold to be missing

Configuration file

Name	*Options*	*Explanation*
LogTransform	"true"/"false"	Perform Log10 transformation on data before correction (to ensure normal distribution)
RemoveZeros	"true"/"false"	There is no such thing as zero. Samples is either detected (value) or not detected (missing). Zeros mess up statistics.
OutlierScope	"Local","Global", "Global&Local"	Perform outlier detection per batch, across all the batches, or both
OutlierMethod	"None","Percentile","Linear","Quadratic","Cubic"	The 'linear'/quadratic/cubic' options fit a simple polynomial function to the QC data using robust least squares regression. Points outside the population confidence interval are deemed outliers. The 'percentile' option implements the standard non-parametric < Q1 − 1.5 IQR or > Q3 + 1.5 IQR outlier detection method on the QC samples.
OutlierCI	value between between 0.9 and 0.99	Set the Confidence interval for the ploynomial methods above. 90%-99%
OutlierPostHoc	"Ignore","MPV","NaN"	For the corrected data. Either ignore outliers or replace with the QC median peak value (MPV) or remove (replace with "not a number" NaN).
IntraBatchMode	"Median","Linear","Spline"	Three correction modes. "Spline" is the default QCRSC algorithm that requires optimisation of the smoothing parameter. "Linear" is a simple (interpolated) linear regression based on the QC values & as such requires no smoothing optimisation. "Median" equalises the QC median across batches & ignores within batch systematic change.
InterBatchMode	"QC","Reference"	QC' = both the intra- and inter-batch based on the QC samples; 'Reference' = intra-batch correction based on the QCs but the inter-batch correction uses the samples labelled as 'Reference'.
QCRSCgammaRange	"x:y:z"	The range of values (x to z) to search to determine to optimal correction curve (in increments of y). A value >=4 results in a curve equivalent to a linear regression. A value < 0 results in a highly nonlinear curve.
QCRSCcvMethod	"3-Fold","5-Fold","7-Fold","Leaveout"	Type of cross-validation used for optimising the smoothing parameter value.
QCRSCmcReps	integer	The number of Monte Carlo (random) resamples of the k-fold cross validation. The resulting cvMSE the mean of the generated set of cvMSEs.
CorrectionType	"Subtract","Divide"	Subtract or divide the correction curve from the raw data. Optional depending on whether you believe the bias to be additive or multiplicative. Recommendation: If multiplicative & Log10 transformation is used then Subtract.
BlankRatioMethod	"QC","Median","Percentile"	BlankRatio = 100*max(BlankValue)/SampleReferenceValue. 'QC' sets the SampleReferenceValue as the median QC value; 'Median' sets the SampleReferenceValue as the median Sample value; 'Percentile' sets BlankRatio = % of Samples < max(BlankValue)*RelativeLOD (RelativeLOD is a user defined constant - default 1.5). In this instance missing values are considered Blank values
RelativeLOD	value	The RelativeLOD is a multiplier (relative to the Blank) used to set a value below which a peak is considered background noise. i.e., "estimated LOD = max(Blank) x relativeLOD"
StatsParametric	"true"/"false"	Choose whether summary statistics for each corrected feature are calculated using parametric (mean, standard deviation etc) or non-parametric methods (median, median absolute deviation etc).
ParallelProcess	"true"/"false"	Switch on the parallel processor to use mulitple processor cores to speed up the QCRSC engine.

Clean & Explore

Name	***Explanation***
Filter:	
batch-wise/complete	Should the filter bank calculate its statistics across all batches or calulate each batch individually and select the worst batch to represent that peak.
% missing thesh =	% allowed missing samples. e.g. if 20% then peak is removed if number of missing > 20%.
Before/After	Compare the performance statistics/visualisations before or after correction. Use this to convince yourself of the utility of the correction algorithm.
PCA preprocessing:	
log2(Fold wrt QC)	First perform log2 fold change - with fold = sample value / mean QC value.
log10 Transform	*OR* log10 transformation
Autoscale / Pareto scale	Preform autoscaling (mean centre then divide by standard deviation) or Pareto Scaling (mean centre then divide by sqrt(standard deviation). Scaling is always performed after the transformation.
PCA Missing value imputation:	
KNN column	KNN missing value imputation replacing with the nearest metabolite feature
KNN row	KNN missing value imputation replacing with the nearest sample
k =	replaces missing values with a weighted mean of the k nearest-neighbor columns/rows. KNN imputation always performed after transformation & scaling.
blank/20%min	replaces missing values with the maximum blank value for that peak or if no blanks detected 20% of the lowest value. blank/20%min is always performed before transformation or scaling.
PCA projection:	
Project in QC samples	PCA model is generated using only the Sample data. This removes any possible bias from the QC, Blank, or Reference samples. The QC, Blank, or Reference sample data can applied to the PCA model (projected through) and plotted with the Sample data.
Project in Blank samples	
Project in Reference samples	