Cheat Sheet

Feature Inclusion Prefilter

Option	*Action*
% QCs missing per feature <=	Feature is kept if there are less than (or equal to) the specified % of missing QC samples.
↓ Every Batch	% missing QC calculated per batch and peak kept if every batch has <= specified % missing QCs
↓ Any Batch	% missing QC calculated per batch and peak kept if any batch has <= specified % missing QCs
↓ Complete	% missing QC calculated across all batches
☑ QC < LOQ* = missing	consider any QC sample with signal < the LOQ* to be missing
LOQ* = ☐ x max(Blank)	estimated Limit Of Quantification = ☐ x maximum Blank value

QC Correction Parameters

Label	*Options*	*Explanation*
Log Transformed Correction	"true"/"false"	Perform Log_{10} transformation on data before correction (to ensure normal distribution)
Remove Zeros	"true"/"false"	There is no such thing as zero. Samples is either detected (value) or not detected (missing). Zeros mess up statistics.
QC Outlier Detection Method	"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.
QC Outlier Detection %CI	value between between 0.9 and 0.99	Set the Confidence interval for the ploynomial methods above. 90%-99%
QC Outlier Replacement Strategy	"Ignore","Median","NaN"	For the corrected data. Either ignore outliers or replace with the QC median value or remove (replace with "not a number" NaN).
Within-Batch Correction Mode	"Mean","Linear","Spline","Sample"	Three correction modes. "Spline" is the default QCRSC algorithm that requires optimisation of the smoothing parameter. "Linear" is a simple Robust (bisquare) linear regression based on the QC values & requires no smoothing optimisation. "Mean" equalises the QC mean across batches & ignores within batch systematic change. "Sample" ignores the QCs and corrects using linear regression based on the samples labelled 'Sample'. N.B. If the number of QC samples in a batch is 5 or less then the Spline Algorithm switches to a linear correction (N<=5 is too few for effectively fitting a non-linear curve with cross-validation). N.B. The 'Sample' option is not recommended and is primarily included to illustrate how poor this approach is to within-batch correction (or as a last resort for data with no QCs).
Between-Batch Correction Mode	"QC","Reference","Sample"	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'. 'Sample' = both the intra-batch & inter-batch correction uses the samples labelled as 'Sample'. N.B. The 'Sample' option is not recommended and is primarily included to illustrate how poor this approach is to between-batch correction (or as a last resort for data with no QCs).
QCRSC Cross Validation Method	"3-Fold","5-Fold","7-Fold","Leaveout"	Type of cross-validation used for optimising the smoothing parameter value.
QCRSC Monte Carlo repetitions	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.
Blank Ratio Method	"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

Clean & Explore

Name	***Explanation***
Missing Filter:	Remove features where there are high number of missing values.
Every-Batch/Any-Batch/Complete	Set whether the filter bank calculates its feature-wise statistics across all batches (complete) or calulates for each batch individually. If 'Every Batch' then every batch must pass the % missing threshold for a given feature to be retained, else if 'Any Batch' then only one batch as to pass the threshold for the feature to be kept.
% missing per feature <=	% allowed missing Samples. e.g. if 20% then a given feature is removed if number of missing samples labeled 'Sample' > 20%.
Before/After	Compare the performance statistics/visualisations before or after correction. Use this to convince yourself of the utility of the correction algorithm.
Filter Bank:	
ON/OFF	Apply the dial setting to the feature selection filter.
Mode	Leave on 'Complete' unless performing large scale studies over many batches. For a given feature: 'Max qcRSD' deferes to the using the stats from the worst batch (batch with highest qcRSD) for all filter settings. 'minQC' deferes to the best batch (batch with lowest qcRSD) for all filter settings. 'Median qcRSD' deferes to the most representative batch for all filter settings. 'Complete' calulates the stats across all batches.
PCA preprocessing:	
log10 Transform	apply a log10 transformation to all data.
Autoscale / Mean Center	Perform autoscaling (mean center then divide by standard deviation) or just mean center to each individual feature. Scaling is always performed after the transformation.
PCA Missing value imputation:	
KNN column	KNN missing value imputation replacing with the nearest 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 feature 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	The PCA model is created 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	