Data normalization options

cellHTS2 provides different options for handle and normalize of high-throughput screening data. We here provide a short overview of normalization strategies.

Introduction

cellHTS2 implements a number of different normalization options to scale plate-to-plate differences in an experiments. In general, one distinguished between control-based and sample-based normalization methods that have both advantages and disadvantages. Sample-based normalization strategies are mostly used if it assumed that the number of "hits" in a plate is rather low, and fails to be robust if many wells show phenotypic changes. This is particularly important to consider, if RNAi (or compound) reagents have a non-random distribution across the experiments, such as the case of several siRNA libraries and in experiments that are designed to retest previously identified "hits". Control-based normalization methods can avoid such pitfalls, however, since the number of controls per plate is usually limited, are prone to variations in the control wells. Spatial effects can be corrected by B-score and Loess transformation.

In general, the choice of normalization method is very much dependent on the experiment design and the quality of the experimental results. There is no "general" recommendation that can be given for all experiments, but we usually advice to start with simple (e.g. median) normalization methods, assess the data by plate plots and check whether spatial normalization might be necessary.

Median normalization (sample-based normalization)

Plate median normalization scales plates based on calculating the relative signal of each well compared to the median of all sample wells in the plate. The median is calculated of the sample wells (e.g. for all wells that contain reagents that target genes of interest) in a given result file. In many cases, plate normalization is the preferred (and least stringent) normalization option but should be avoided if there are spatial effects on plates in the HTS data set.

Shorth normalization (sample-based normalization)

Shorth normalization is a variant of the plate median normalization which consist of using the midpoint the shorth of the per-plate distribution of values on sample wells. This is for example appropriate, if distribution of sample values due to non-random distribution of reagents throughout an experiment has multiple peaks.

Mean normalization (sample-based normalization)

Mean normalization is a variant of the plate median scaling which divides each sample value by the per-plate average. This normalization is less robust against outliers than Median normalization.

Normalization on Negative Controls (control-based normalization)

This method consists of scaling the sample measurements by the per-plate median of the values of the wells that have been annotated in the plate configuration file as "negative controls". This method is particular appropriate if many sample values are likely giving an effect. It should also be noted that the number of negative controls per assay plate should be sufficiently high, otherwise even small differences in negative control values might lead to significant shift in the reported results.

Percent of Control (control-based normalization)

Same as above, except that wells annotated as "positive controls" are used as a reference.

Normalized Percent of Control (control-based normalization)

Normalization methods also known as "normalized percentage inhibition" (NPI) that relies on calculating a well result by dividing the difference between sample measurements and the average of positive controls through the difference between positive and negative controls.

B-Score Normalization

B-score normalization can remove row and column biases within each plate by fitting a two-way median polish to the raw data in a per plate manner. This method is particularly useful if row or column effect in plates are observed (e.g. pipetting differences or evaporation).

Loess Regression and Robust Local Fit Regression

cellHTS2 provides additional spatial normalization methods that fit a polynomial surface to the intensities within each assay plate using local regression and that can be performed via normalizePlates or spatialNormalization functions, although we advise to apply these methods using the former function. The fit can be performed either using the loess procedure or the locfit.robust function of package locfit. In normalizePlates, if method="locfit", spatial effects are removed by fitting a bivariate local regression to each plate and replicate, while if method="loess", a loess curve is fitted instead.

References and further reading

1. Bioconductor/R cellHTS2 description [link]

2. Boutros M, Brás LP, Huber W. (2006). Analysis of cell-based RNAi screens. Genome Biol. 7:R66. [link]

3. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. (2006). Statistical practice in high-throughput screening data analysis. Nat Biotechnol. 24:167-75. [link]

4. Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, Santoyo-Lopez J, Dunican DJ, Long A, Kelleher D, Smith Q, Beijersbergen RL, Ghazal P, Shamu CE. (2009). Statistical methods for analysis of high-throughput RNA interference screens. Nat Methods 6:569-75. [link]