Data normalization in Genevestigator®

Status: September 2013

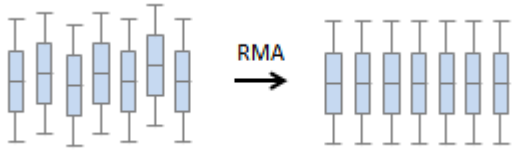
Background and requirements

Genevestigator does not consist of a mere collection of curated studies that are analyzed individually. The essence of Genevestigator is to integrate all studies and perform meta-analysis across thousands of experiments. We therefore looked for a method delivering:

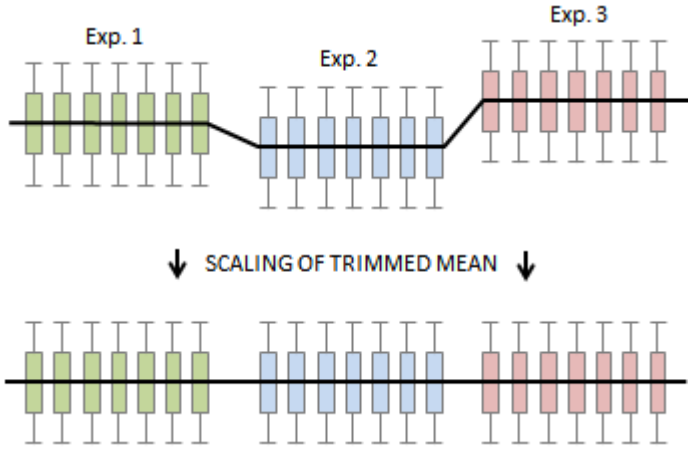
1. **Robust values for comparisons within individual experiments**.
    - precise differential expression analysis between groups of experimental variables
    - compare absolute expression values between samples
2. **Approximate values for comparisons between experiments**.
    - aggregate data across all experiments to generate meta-profiles for anatomy or cancer
    - roughly compare absolute expression values across different studies to find those with extreme expression values (+/- 5% is acceptable for this purpose)

Normalization methods

STEP 1: raw data is processed using **quantile normalization** as implemented for RMA in Bioconductor.



STEP 2: **global scaling** of the trimmed mean (10% trim of all values of a chosen experiment) to a common target value.



Proof of principle

STEP 1 is well accepted in the research community and its value has been shown and discussed extensively in the literature.

STEP 2 was established by Nebion to fulfill the requirement of making absolute expression values sufficiently comparable between experiments. A trimming at 10% and the choice of the mean seemed to yield the best biological results. Below, we show two examples of biological validation.

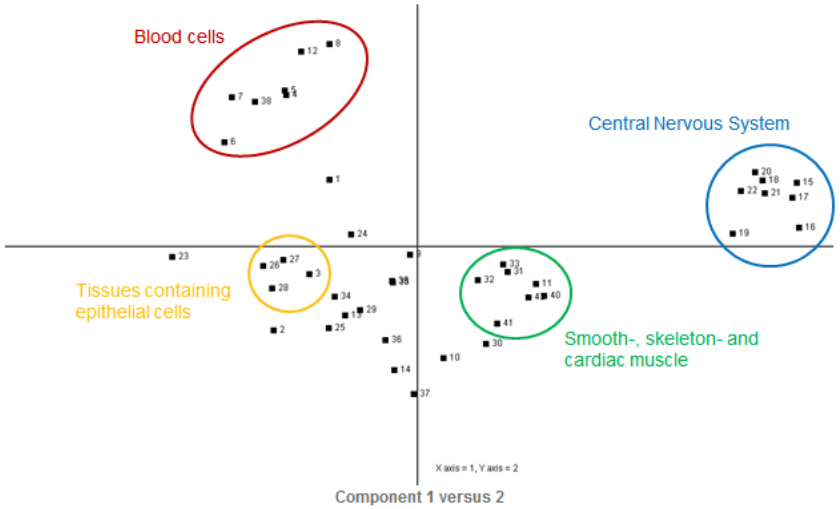
Example 1: Validation of the *Anatomy* meta-profiles

For each tissue type, we created a mean expression vector as aggregated from all samples (arrays) belonging to that tissue type. Typically, expression values from multiple studies and experiments were combined.

We then performed a Principle Component Analysis of all these tissues. The resulting plot for component 1 vs. 2 clearly positioned the tissues into clusters of functionally closely related tissue types.

When overlapping the plots for human, mouse and rat, we obtained highly similar clustering patterns, even though the underlying data were from different origin, labs, and organisms (data not shown; in press).

These results indicate that the *Anatomy* meta-profiles generated in Genevestigator represent, to a very large extent, true biological information rather than lab or batch effects. The same holds true for the profiles displayed in the *Neoplasms* or *Samples* tools.



Example 2: Validation using reference genes

Across a globally normalized database, reference genes for RT-qPCR typically should show stable expression across all datasets. This can easily be visualized in Genevestigator using a panel of such genes. Reversely, the Genevestigator RefGenes tool was developed to identify new reference genes based on globally normalized data. The experimental validation of these genes showed that they perform even better than commonly used genes, indicating that the global normalization performed in Genevestigator is supported by biological evidence. See also Hruz et al., *BMC Genomics* 2011, **12**:156 and www.refgenes.org.