Rat liver microarray data

The file 'JnJ_Codelink_Data.tsv' contains normalized expression data from 617 rat samples run on the GE CodeLink Rat Whole Genome Bioarray. The file is a tsv file with samples as columns, and CodeLink platform features as rows.

Animal information:

All specimens used in the study were 60-70 day old male Sprague-Dawley rats. They were fed "Lab Diet 5001 Rodent Diet" (Supplier: PMI Nutritional International, L.L.C., Brentwood, MO) *ad libitum*. The animals were individually housed, on a 12 h light/dark cycle. Rats were fasted for 24 hrs after last dosing until necropsy the following morning.

Sample information:

The column header provides meta-data about the sample through multiple "|"-separated fields. The fields in the header are:

1. A two-letter *batch identifier*: Samples with the same batch name were run together on the CodeLink platform. Each batch of samples typically includes some vehicle or untreated samples and some treated samples.

2. A *lot identifier*: A batch can include samples from rats obtained from different lots. When comparing a treated sample with a vehicle sample, it is recommended that the comparison be made only with vehicles belonging to the sample batch and lot.

3. An *animal identifier*: This identifier is unique across all samples

4. *Compound* name: The name is either a paradigm compound or 'Vehicle' or 'Untreated'. A short description of each of the 123 paradigm compounds present in this dataset is in the file 'Paradigm_Compound_Description.txt'.

5. *Dosage* in mg/kg. The dosage information field is empty for Vehicle and Untreated samples. The vehicle used was 0.5% methocel administered orally.

6. *Duration* in days: The number in this field indicates how many days the dose was administered. 1d means the dose was administered once, and the rat was sacrificed 24 hours later, 4d means the dose was administered on 4 consecutive days and the rat was sacrificed on the 5^{th} day.

7. *Route of administration*: Indicates if the compound is administered orally or by IP.

For instance, the column with header “AC|b5|AC66|Erythromycin estolate|1500|1d|Oral” can be interpreted as data from rat AC66 which was administered 1500 mg/kg of Erythromycin estolate orally on day 1 and sacrificed 24 hours later. The sample is part of the AC set of samples and the vehicle samples that it can be compared to will have headers of the form “AC|b5|<id>|Vehicle||..”.

Feature information:

The first column of the data file contains CodeLink platform feature identifiers and the second column has probe names. All other columns contain sample data. The ‘Original_Codelink_Annotations.txt’ file contains descriptions of the features. Each feature is associated with a ProbeName , a ProbeType and some additional annotation columns which may be outdated. Overall there are:

- DISCOVERY: 33,489
- FIDUCIAL: 512
- OTHER: 272
- NEGATIVE: 256
- POSITIVE: 240

Each DISCOVERY type feature maps to a unique ProbeName and it is recommended that users only use these probe types in analysis.

The fasta file, ‘CodeLink.RWG.probe.fa’, contains the 30 bp sequence of the probes. The fasta file was used as the basis for re-creating up-to-date annotations for the probes. The pipeline used for this annotation and the resulting classification of probes into 5 different confidence classes are described in the file ‘codelink_probe_mapping_report.pdf’ present in the annotations directory.

Normalization process:

The data file ‘JnJ_Codelink_Data.tsv’ contains log fold change values. These values were obtained batch-wise as follows:

- Each batch includes some vehicle/control samples.
- Data was converted to log scale and quantile normalized.
- Median of control samples was computed (ignoring outliers)
- This median was subtracted from all samples in the batch to obtain log fold change values.