To reproduce the results found on TESLA paper in order to check the performance of the algorithm, understand it and know the effects of the tuning parameters and posterior comparison with other algorithms, we need to gather the data (Drosophila melanogaster cycle time series, ref 1) and preprocess it. To do so we need to follow the same steps than in TESLA paper.

* Get form the data a subset of 4,028 sequence-verified, unique genes.
* Normalization (remove dye-dependent intensive response and the average ratio of signals from the experimental and reference sample equals one). Final expression value is the log ratio of signals.
* Missing values filled by linear interpolation using values from adjacent time points (like ref 2) since we assume that gene expression values change in a smooth way over time. If starting or end point, substitute it by the nearest value.

**DATA RETRIEVAL**

Data: <http://flygenome.yale.edu/Lifecycle> (seems the link is broken) <http://flybase.org/>

Data source: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4347> GEO

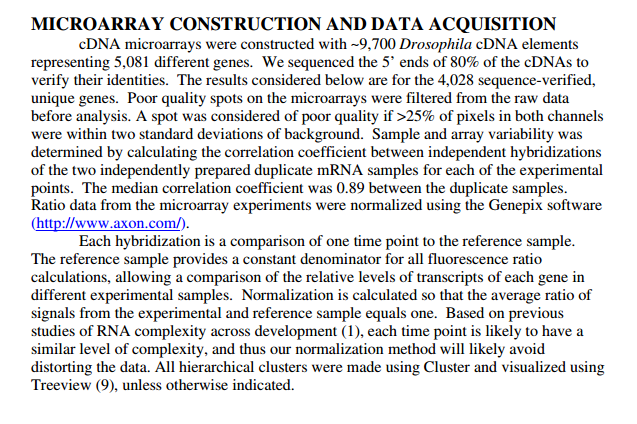
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2837>

**LOADING DATA**

First of all load the data to R. To do so after having a look to the txt files I realized the data is from row 72 in all 4 files (before metadata). After leading and some cleaning of the original files we end with 4 tables where:

* Columns are time steps (not ordered yet, but the headers, that encode the time points, are ordered by name: GSM73861 🡪 18-19h, GSM73862 🡪 19-20h …
* Rows are the different Genes (>9000, we need to trim as done in TESLA paper, ref? in order to leave 4028)

**DATA CLEANING**



**DATA PREPROCESSING**

**TODO:**

* Merge datasets (outer/inner junction, since they have different nrows, doesn’t matter since we will trim this).
* Check which (ordered) columns correspond to which time steps and find an explanation why we have 174 time points and in TESLA the say 66

First: GSM73861 🡪 18-19h Embryo N=2

Last: GSM74030 🡪 Eya Female Adult 0-24hr Sample A PolyA RNA

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4347>

* Chose the same subset of 4028 genes as in TESLA paper (I think is the same subset they use in [1]).
* Normalization by rows????? 🡪 since they represent the SAME gene for different time points…

(So, normalization by genes or by time??) Before or after subset???: I would say after

REFERENCES

1. Arbeitman M, et al. (2002) Gene expression during the life cycle of Drosophila melanogaster. Science 297:2270–2275
2. Zhao W, Serpedin E, Dougherty E (2006) Inferring gene regulatory networks from time series data using the minimum description length principle. Bioinformatics 22(17):2129–2135.