A pipeline to analyse time-course gene expression data

true
true

Abstract

The phenotypic diversity of cells is governed by a complex equilibrium between their genetic identity and their environmental interactions: Understanding the dynamics of gene expression is a fundamental question of biology. However, analysing time-course transcriptomic data raises unique challenging statistical and computational questions, requiring the development of novel methods and software. Using as case study time-course transcriptomics data from mice exposed to different strains of influenza, this workflows provides a step-by-step tutorial of the methodology used to analyse time-course data: (1) normalization of the micro-array dataset; (2) differential expression analysis using functional data analysis; (3) clustering fo time-course data; (4) interpreting clusters with GO term and KEGG pathway enrichment analysis.

```
## Warning: no function found corresponding to methods exports from 'XVector'
## for: 'concatenateObjects'
```

TODO list

- Add legends to the plots that don't have.
- Currently have two functions "plot_genes", "plot_centroids", that do exactly the same thing. Find a better name for both of them

Introduction

Gene expression studies provide simultaneous quantification of the level of mRNA from all genes in a sample. High-throughput studies of gene expression have a long history, starting with microarray technologies in the 1990s through to single-cell technologies. While many expression studies are designed to compare the gene expression in distinct groups, there is also a long history of time-course expression studies, where the the gene expression is compared across time by measuring mRNA levels from different samples across time.[1] Such time course studies can vary from measuring a few distinct time points, to sampling ten to twenty time points. Many longer time series are particularly interested in investigating development over time. More recently, single-cell studies track single cells through their development, and a single cell is measured at a particular moment in its developmental progression – a value that is not know but estimated from the data as its "pseudo-time."

While there are many methods that have been proposed for discrete aspects of time course data, the entire workflow for analysis of such data remains difficult, particularly for long, developmental time series. Most methods proposed for time course data are concerned with detecting genes that are changing over time (differential expression analysis), examples being `edge` (Storey et al. 2005), functional component analysis based models (Wu and Wu 2013), time-course permutation tests (Park et al. 2003), and multiple testing strategies to combine singe time point differential expression analysis (Wenguang and Zhi 2011). However, with long time course datasets, particularly in developmental systems, a massive number of genes will show *some* change. For example, in the mice lung tissues infected with influenza, over 50% of genes are shown to be changing over time. The task in these settings is often not to detect changes in genes, but to categorize into biologically interpretable patterns the vast number of changes discovered.

We present here a workflow for such an analysis that consists of 4 main parts (Figure **??**):

- Quality control and normalization;
- Identification of genes that are differentially expressed;

[1] Because the collection of the mRNA is often destructive, samples at different time points are generally from different biological samples; longitudinal studies, for example tracking the same subject over time, are certainly possible in certain settings, but not directly considered here.

- Clustering of genes into distinct temporal patterns;
- Biological interpretation of the clusters.

This workflow represents an integration of both novel implementations of previously established methods and new methodologies for the settings of developmental time series. It relies on several standard packages for analysing gene expression data, some specific for time-course data, others broadly used by the community. We provide the various steps of the workflow as functions in a R package called `moanin`.

Park, Taesung, Dong-Hyun Yoo, Jun-Ik Ahn, Seung Yeoun Lee, Seungmook Lee, Sung-Gon Yi, and Yong-Sung Lee. 2003. "Statistical tests for identifying differentially expressed genes in time-course microarray experiments." *Bioinformatics* 19 (6): 694–703. https://doi.org/10.1093/bioinformatics/btg068.

Storey, J. D., W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. 2005. "Significance analysis of time course microarray experiments." *Proc. Natl. Acad. Sci. U.S.A.* 102 (36): 12837–42.

Wenguang, Sun, and Wei Zhi. 2011. "Multiple Testing for Pattern Identification, with Applications to Microarray Time-Course Experiments." *Journal of the American Statistical Association* 106 (493). Taylor & Francis: 73–88. https://doi.org/10.1198/jasa.2011.ap09587.

Wu, Shuang, and Hulin Wu. 2013. "More Powerful Significant Testing for Time Course Gene Expression Data Using Functional Principal Component Analysis Approaches." *BMC Bioinformatics* 14 (1): 6. https://doi.org/10.1186/1471-2105-14-6.

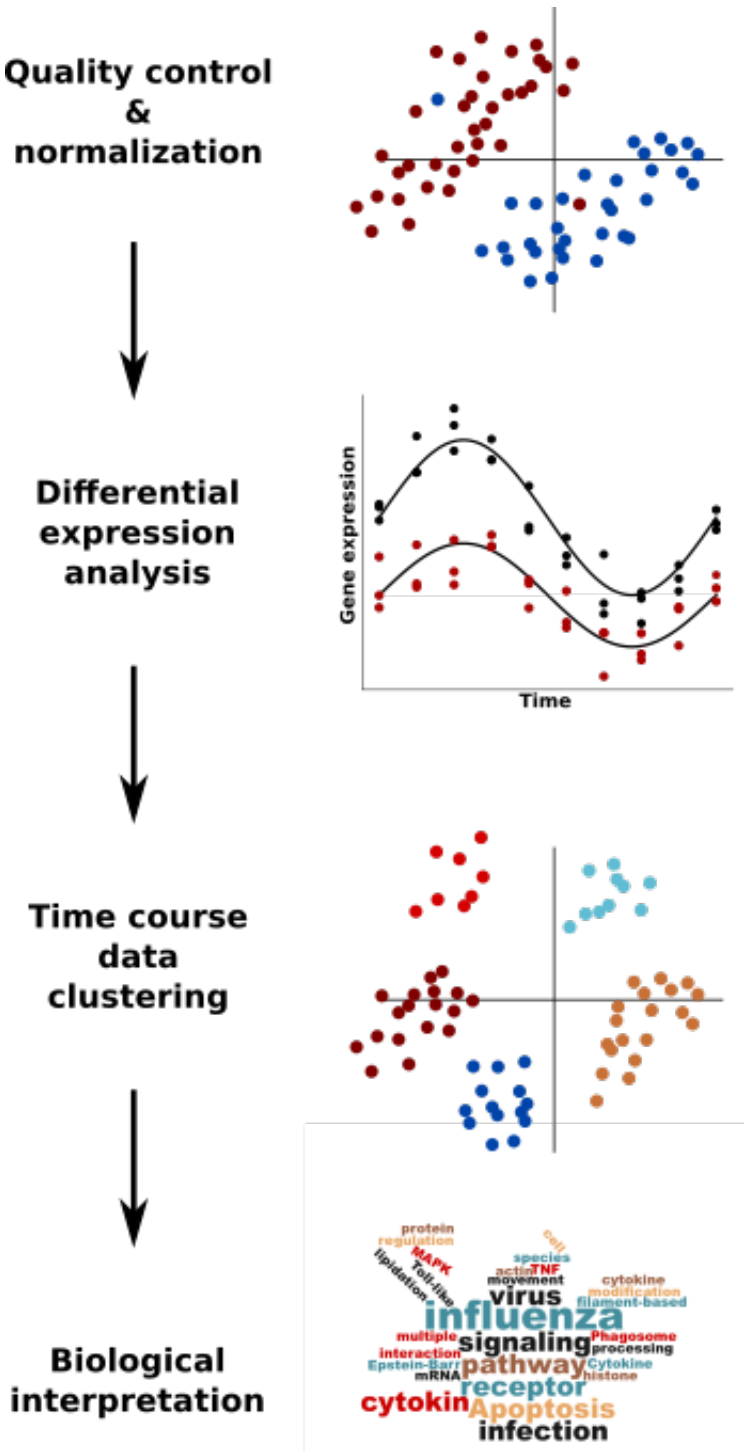

Figure 1: Workflow for analyzing time-course datasets.