PAGODA: pathway and gene set overdispersion analysis characterizes single cell transcriptional heterogeneity

Jean Fan[1,2], Neeraj Salathia[3], Rui Liu[4], Gwen Kaeser[5], Yun Yung[5], Joseph Herman[1], Fiona Kaper[3], Jian-Bing Fan[3], Kun Zhang[4], Jerold Chun[5], Peter V. Kharchenko[1]

[1]Center for Biomedical Informatics, Harvard University, Boston, MA, USA; [2]Bioinformatics and Integrative Genomics Program, Harvard University, Boston, MA, USA; [3]Illumina Inc., San Diego, CA, USA; [4]Department of Bioengineering, University of California, San Diego, CA, USA; [5]Dorris Neuroscience Center, Molecular and Cellular Neuroscience Department, The Scripps Research Institute, La Jolla, CA, USA

ABSTRACT

The transcriptional state of a cell reflects a variety of biological factors, from cell-type-specific features to transient processes such as the cell cycle, all of which may be of interest. However, identifying such aspects from noisy single-cell RNA-seq data remains challenging. We developed pathway and gene set overdispersion analysis (PAGODA) to resolve multiple, potentially overlapping aspects of transcriptional heterogeneity by testing gene sets for coordinated variability among measured cells.

SOFTWARE AVAILABILITY

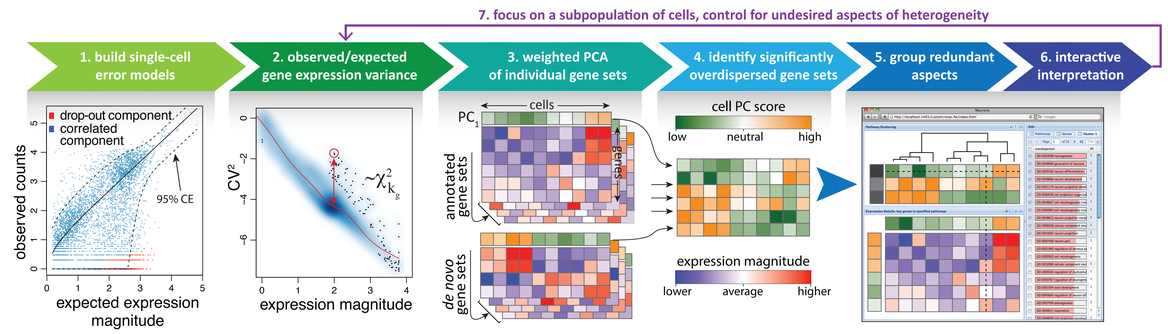
The PAGODA functions are implemented in version 1.99 of the scde R package, available at **pklab.med.harvard.edu/scde/** along with additional tutorials and resources, as well as on BioConductor. Source code and development versions are also available on Github at **github.com/hms-dbmi/scde**. Requires R >= v3.0.0.

FIND US ON

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

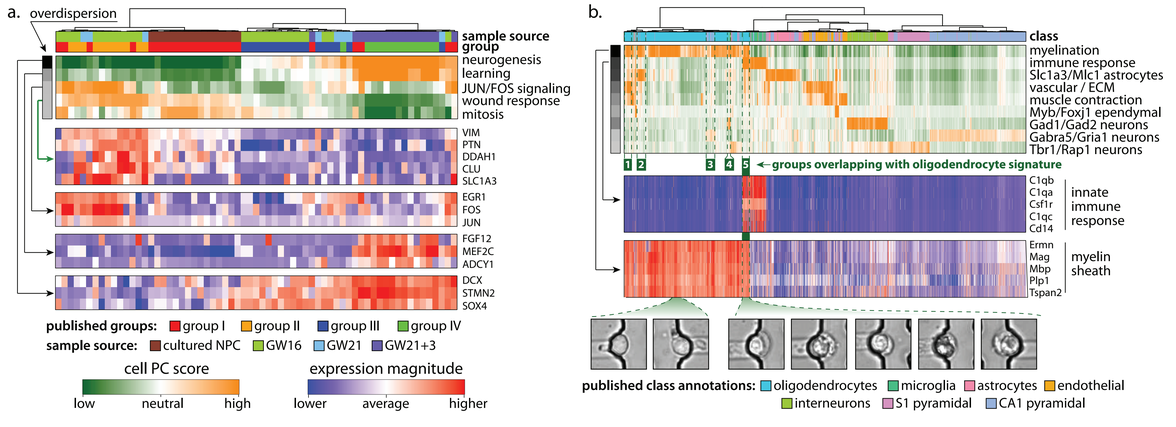
github
SOCIAL CODING

Overview of Pathway and Gene set Overdispersion Analysis (PAGODA)



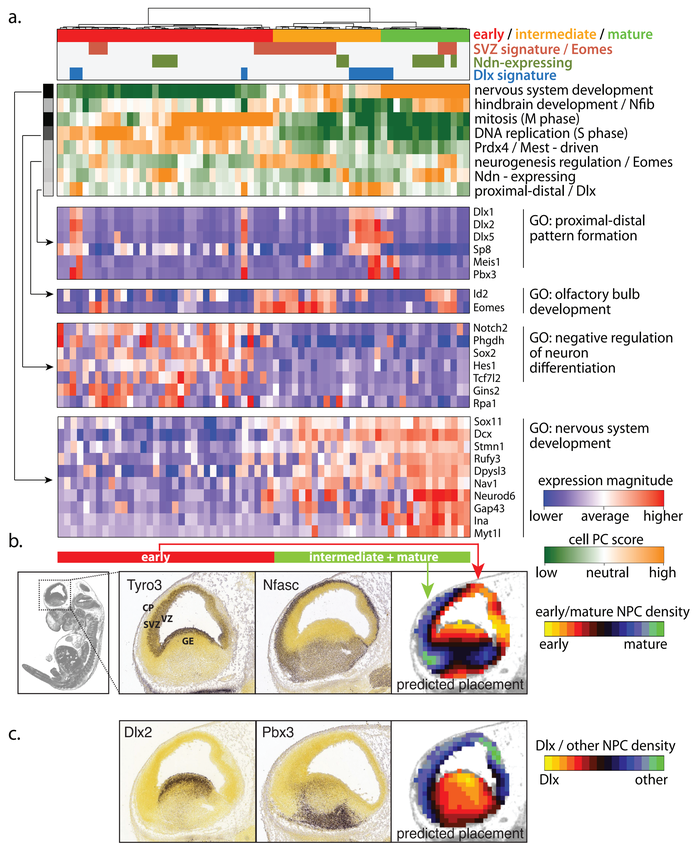
1. Cell error model fitting to quantify the dependency of amplification noise and drop-out probabilities on the expression magnitude; **2.** Variance normalization, taking into account the uncertainty in the variance estimates of each gene by determining effective degrees of freedom; **3.** Weighted PCA analysis is performed independently on functionally-annotated gene sets, as well as de novo gene sets determined based on correlated expression in the current dataset; **4.** If the amount of variance explained by a principal component of a given gene set is significantly higher than expected, the gene set is called overdispersed, and the cell scores defined by that principal component are included as one of the significant aspects of heterogeneity in the dataset; **5.** Redundant aspects are grouped; **6.** PAGODA implements a web browser-based interface; **7.** Iterate if needed

PAGODA recapitulates published findings and identifies additional aspects of heterogeneity



The dendrogram and heatmap PAGODA results. Column colors indicates group assignments from the original published analysis. **a)** Transcriptional heterogeneity in mixture of 65 cultured human neuronal projenitor cells and primary cortical cells from *Pollen et al.* An additional "wound response" aspect, which describes a pattern complementary to the activation of neurogenesis pathway, can be seen (green arrow). **b)** Transcriptional heterogeneity in 3005 mouse cortex and hippocampus cells from *Zeisel et al.* PAGODA identifies microfluidic traps with multiple transcriptional signatures, suggesting capture of two associated cells of different types.

PAGODA characterizes transcriptional heterogeneity of NPCs in embryonic mouse cortex



(right) PAGODA clustering identified previously unassessed genes that are predicted to identify distinct populations of NPCs; **a)** scRNA-seq expression magnitudes of the two selected genes relative previous cell clustering and classification; **b-g)** Coronal E13.5 brain sections labeled by in situ hybridization probes for Rpa1, found in proliferating clusters (b-d) and Ndn in more mature clusters (e-g), representing genes with previously unknown relationship to NPCs; **b-d)** Rpa1 shows high expression in the VZ and SVZ with reduced expression in the region of young postmitoic neurons located in the superficial CP; **e-g)** Ndn expression is promient throughout the CP. There are also rarer high expressing cells in the VZ and SVZ (black arrows) consistent with scRNA-seq data.

(left) Transcriptional diversity of neuronal progenitor cells in the developing mouse brain; **a)** Eight significant aspects that were detected are labeled by their primary GO category or driving genes. Color codes in the top panel summarize key subpopulations of NPCs distinguished by the detected heterogeneity aspects; **b)** Anatomical placement of the early vs. maturing NPC classes within embryonic brain. Computational prediction of spatial distribution of early vs. maturing NPCs based on the overall transcriptional profile (third panel) places early NPCs near VZ, and maturing ones in SVZ/CP regions, consistent with known placement of apical (early) and basal (intermediate) progenitors; **c)** Anatomical placement of the Dlx-expressing NPCs. Computational prediction places such cells in the GE (ganglionic eminence region), consistent with the anatomical origination of the tangentially-migrating NPCs.