## Defining Gene Regulatory Networks During Corticogenesis

Ryung Kim,[1] Robert Bachoo,[2] Wing Wong,[3,4] Ronald DePinho[5]

Abstract: While the information revolution has made great impact on our ability to integrate gene sequences, gene ontology, and protein structure information, it has proven much more challenging to combine this with large scale gene expression data. The development of the mammalian brain involves sequential modifications in gene expression regulated by a cascade of transcription factors. Genome scale analytical tools have yet to be developed which are capable of identifying critical transcriptional regulators underlying lineage specification in the developing telencephalon. We generated a large expression data set of the developing mouse brain from embryonic day 8 to postnatal day 10.

A resampling based clustering algorithm was developed to extract stable gene clusters with strong co-membership without forcing all genes into clusters. The algorithm uses average co-membership matrix but also requires each cluster to maintain quality in terms of original feature distances. We showed that our algorithm performs better than traditional clustering algorithms as well as other existing resampling based methods on several simulated data sets as well as on microarray data sets. We applied our new clustering algorithm on the cortex development data to obtain 45 stable clusters, which could be distinguished into 4 basic patterns: down regulating, transiently up-regulating with peak expression corresponding to mid-gestation, up-regulating with peak in early postnatal, and up-regulating with peak in young adulthood.

In an effort to find common known cis-regulatory elements that are shared in the promoter regions of the co-expressed genes, we have searched the upstream regions of co-expressed genes: we mapped 147 known mouse transcription factor binding site motifs at promoter regions of all genes in the mouse mRNA chip (known binding motifs are obtained from public Transfac database) . This mapping result enabled us to find candidate TF's for each group that is identified by the expression data. We propose a novel approach to use binding motif information and expression information to *simultaneously* search for the co-regulated subset of genes and the corresponding transcription factors in a co-expressed gene group. We applied our novel strategy to the astrocyte specific expression clusters to define 3 candidate co-regulated genes as those genes which share 7 common transcription factor binding motifs in their promoter region. We identified three signature genes, glial fibrillary acidic protein, and Aquaporin 4, and a novel gene. We identified a SRY-box-containing factor as a particularly promising candidate as the transcription factor. In-situ hybridization and immunohistochemical analysis has confirmed the temporal and spatial expression pattern of this transcription factor. We now have conclusive biological validation that this transcription factor is critical for regulating cell fate lineage decisions of undifferentiated neural stem cells in the mouse brain.

Key words: Corticogenesis, brain development, lineage specification, neural stem cell, microarray, gene expression, transcription factor, Trasncription Factor binding motifs, Resampling based clustering, clustering.

**ACKNOWLEDGEMENTS.** This research was supported in part by NIH grants HG-HG02341 and P01 CA095616.

[1,] Mathematical Sciences Dept, Worcester Polytechnic Institute, Worcester, MA, USA

[2] Departments of Neurology and Medicine, University of Texas Southwestern Medical Center, Dallas, Texas, USA

[3] Department of Statistics, Stanford University, Stanford, CA, USA

[4] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA