**Self-organizing maps for analysis and mining of functional genomic data**

Ali Mortazavi[1*], Shirley Pepke[2*], Georgi Marinov[1], Barbara Wold[1]

[1]Biology Division, California Institute of Technology, Pasadena, CA

[2]Center for Advanced Computing Research, California Institute of Technology, Pasadena, CA

* These authors contributed equally to this work

## Abstract

The analysis and mining of multiple genome-wide datasets in an integrated way is a fundamental challenge in functional genomics. The Self-Organizing Map (SOM) is a widely applied unsupervised, machine-learning method used in clustering that is particularly suited for parsing similarity between elements at both global and local levels while coping with very high dimensional data and its associated combinatorial complexity. These properties make it possible to recover diverse correlations and relationships in the input data as patterns on the map after training. We used SOMs to analyze different segmentations of a previously published dataset of 41 human ChIP-seq experiments performed on CD4 T-cells. We present gene-level interpretability of the SOM by using Gene Ontology as a criterion for choosing the optimal map size and further used it to compare different genome segmentations as a function of map size. We show that the maps can be used to understand the global behaviors of histone modifications as well as to identify specific DNA segments from functionally related genes that show similar mark combinations. Additional data-types not used in the SOM training can also be placed on the map, alone or subtractively, to identify further relationships with the histone mark code. Meta-clustering of the map produces functional

states analogous to those produced in a prior study using Hidden Markov Models (HMM) on the same datasets. Mapping the HMM states onto the SOM shows how mixtures of units explain the presence of previously described HMM states that frequently transition amongst themselves or have common functional assignments. Genomic SOMs are a promising additional tool for probing genome-scale structure and function.