

FLOCK: a density-based clustering method for automated identification and comparison of cell populations in high-dimensional flow cytometry data

Yu Qian^{1,2}, John Campbell³, Megan Kong², Yue Liu³, and Richard H. Scheuermann^{*1,2}

¹Division of Biomedical Informatics, ²Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

³Health Information Systems, Northrop Grumman, Inc., Rockville, MD 20850, USA

FLOCK (Flow Clustering without K) is an automated software system developed for the analysis of high-dimensional flow cytometry (FCM) data [Qian 2010]. Unlike traditional model-based approaches, FLOCK employs a grid-based partitioning and merging scheme to identify density-based data clusters. The number of clusters is decided based on the density gap between partitioned data regions in different subspaces. FLOCK has been applied to many FCM datasets, including all five FlowCAP datasets in all four types of challenges with encouraging results. The grid-based approach makes the system highly efficient in identifying dense regions in very large data sets. The use of dimension selection and normalization makes the system robust for the analysis of datasets with different number of markers. FLOCK has been used to identify both known and novel cell populations (see [Qian 2010] for 17 B-cell populations identified by FLOCK). Population statistics are also calculated, including mean fluorescence intensity (MFI), proportions, coefficient of variation (CV), and expression profiles of each population. Populations identified by FLOCK can be mapped based on their centroid positions and compared across samples that use the same reagent panel. FLOCK has been implemented in the publically available Immunology Database and Analysis Portal - ImmPort (<http://www.immport.org>) with an advanced graphical user interface and visualization for open use by the immunology research community, where we are also in the process of linking FLOCK results to cell types defined in the Cell Ontology (CL) [Diehl 2010] for population interpretation and knowledge integration. We have also developed our own FCS conversion and transformation method that generates output consistent with FlowJo (TreeStar, Inc.), to ultimately facilitate an end-to-end free analysis pipeline that starts from instrument FCS files to biologically interpreted cell populations.

Supported by U.S. NIAID Bioinformatics Integration Support Contract (N01-AI40076)

References:

Qian, Y., Wei, C., Eun-Hyung Lee, F., Campbell, J., Halliley, J., Lee, J. A., Cai, J., Kong, Y. M., Sadat, E., Thomson, E., Dunn, P., Seegmiller, A. C., Karandikar, N. J., Tipton, C. M., Mosmann, T., Sanz, I. and Scheuermann, R. H. (2010), Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry Part B: Clinical Cytometry*, 78B: S69–S82. doi: 10.1002/cyto.b.20554

Diehl AD, Augustine AD, Blake JA, Cowell LG, Gold ES, Gondre-Lewis TA, Masci AM, Meehan TF, Morel PA, Nijnik A, Peters B, Pulendran B, Scheuermann RH, Yao QA, Zand MS, Mungall CJ. Hematopoietic cell types: Prototype for a revised cell ontology. *J Biomed Inform.* 2010 Feb 1. [Epub ahead of print] PubMed PMID: 20123131; PubMed Central PMCID: PMC2892030.