Unsupervised analysis of leukemia and normal hematopoiesis by joint clustering of gene expression data

Leukemia is a very heterogeneous cancer of the hematopoietic system. Since its main cause consists of genomic defects in the hematopoietic stem or progenitor cells and given the high complexity of the hematopoietic system, it may seem an important task to investigate the transcriptomic similarities and differences between leukemia subtypes and hematopoietic cells (stem cells, progenitors and differentiated cells).

In this paper, we integrate the largest publicly available gene expression datasets of leukemia and normal hematopoiesis with the aim of uncovering the main gene modules involved in normal hematopoiesis as well as in the various leukemia subtypes.

Using a joint consensus clustering algorithm, we have been able to relate the major leukemia types to their putative cells of origin in an unsupervised manner. While the normal hematopoietic cell modules are also active in leukemias of the corresponding cell type, our approach has determined leukemia-specific modules comprising genes with a known involvement in leukemogenesis.

The expression modules uncovered implicate an unusually large number of transcription factors. This speaks against very simple models of normal hematopoiesis and leukemogenesis that involve just a handful of critical TFs, arguing for the interplay of complex transcription factor networks, in line with the findings of the FANTOM consortium for leukemia and Novershtern et al. for normal hematopoiesis.