Identifying cell-type marker genes using natural language processing

Single-cell RNA-sequencing allows us to measure gene expression levels in thousands of individual cells from a heterogeneous tissue sample simultaneously. After sequencing, cells with similar expression profiles can be clustered together representing groups of distinct cell types. While clustering the cells is relatively straightforward, deciding which cell type each cluster represents is challenging. Researchers often look for enriched expression of known cell-type marker genes within a cell-type cluster in order to make these assignments, but curating lists of marker genes from the scientific literature is time consuming and can lead to inconsistent marker gene lists from different research groups, hindering reproducibility. We hypothesize that natural language processing (NLP) can be used to identify useful markers for thousands of cell types in an unbiased manner. Given large volumes of text, NLP describes the context of any given word numerically, and words related to the same concept will have similar numeric representations. We employed NLP techniques to “read” millions of Pubmed abstracts and quantified the similarity between numeric representations of all genes and cell types, giving us an NLP similarity score for each gene/cell-type pair. We predicted that genes with high NLP similarity to one cell type, but not to other related cell types, would make good cell-type specific marker genes. Using the Cell Ontology, which contains a record of the hierarchical relationships between thousands of cell types, we grouped related cell types and identified genes more similar to each cell type than the others in the group. We tested the utility of these NLP derived cell-type markers for assigning labels to single-cells and found that our approach identified peripheral blood mononuclear cell subtypes as accurately as hand-curated marker genes. This work provides a proof of principle that NLP approaches can be used to create unbiased lists of cell-type-specific marker genes.