Reducing technical dropout events in single cell RNA-seq through molecular depletion

Genomics

Justin Kermack (justin.k@jumpcodegenomics.com), Jumpcode Genomics

Regardless of method, single cell RNA-seq only captures a small fraction of the transcriptome of each cell. Often, this is due to inherent limitations of the methodology as reads 'dropout' at each step of library preparation. These dropout events are then confounded with noise, outliers, and stochastic genetic variation, resulting in the daunting computational task to parse out the true signal. Almost all computational algorithms have evolved to address this zero-inflation issue through a multitude of approaches, typically through various dimensionality reduction or imputation techniques. While consensus for a standardized computational approach has yet to be met, we present a turnkey molecular solution that drastically reduces dropout events attributable to technical noise, statistically enhancing biological interpretation.

Traditionally, single cell data processing incorporates certain filtering and normalization steps prior to canonical clustering and downstream interpretation. Instead of removing those reads in-silico, our universally incorporated molecular solution removes those reads in-vitro, redistributing sequencing clusters to unique biologically relevant transcripts. Our method, called CRISPRclean, seamlessly incorporates into any single cell RNA-seq workflow prior to the final PCR amplification. CRISPRclean leverages CRISPR Cas9 depletion and a custom designed guide library to remove reads that were previously never incorporated in downstream analysis. For example, by analyzing a cohort of publicly available single cell 10x data from various sources, roughly 30-50% of reads aligned to the genome but not the transcriptome, and thus, are conventionally ignored. By tailoring guides to deplete these genomic intervals in addition to the highest expressed protein coding ribosomal and mitochondrial genes, we have exhibited the ability to redistribute 50% of reads through in-silico depletion across single cell data from 14 tissue types.

As an initial proof of concept, we performed our preliminary single cell depletion on a cohort of patient derived endothelial cells from coronary and pulmonary arterial beds. As a result of depletion, 30% of reads were re-distributed, resulting in a net 26% gain in additional unique molecular identifiers (UMIs) on a per cell basis. By leveraging existing algorithms such as molecular cross validation, affinity-based dispersion methods, and random matrix theory, we demonstrate the power of depletion to reduce technical dropouts, enhance the number of genes contributing to biological signal by 858 genes on average, while simultaneously mitigating any reduction in variation. The added benefit of performing depletion translated into the discovery of a biologically relevant phenotype illuminating the transcriptional disposition of coronary arterial beds toward atherosclerosis.