UMI Deduplication in Partek Flow

Most single cell RNA-seq library prep kits compensate for the small quantity of starting material by PCR amplifying the reverse transcribed cDNA. Because some sequences will amplify preferentially, the final proportions of PCR amplified molecules will diverge from the original number of molecules. To correct for these PCR artifacts, reverse transcribed molecules are tagged with unique nucleotide sequences, termed unique molecular identifiers (UMIs), prior to amplification. These UMIs are retained through PCR amplification, allowing PCR products that were amplified from the same original molecule to be identified. Counting UMIs for each gene instead of reads allows the original number of molecules corresponding to each gene to be more faithfully represented.

Identifying reads with matching UMIs and consolidating them into a single aligned read for use in quantification is handled by the Deduplicate UMIs task in Partek Flow.

Default behavior

An important consideration when analyzing UMI data are the errors introduced into the UMIs themselves during PCR amplification of the original molecule. If these errors are not accounted for and each sequenced UMI is considered to be representative of the original UMI, the number of unique molecules can be significantly overestimated. To account for this, the Deduplicate UMIs task uses an implementation of the UMI-tools algorithm described in Smith et al. 2017. Paired-end read support was further improved by incorporating components of the UMI deduplication tool Connor.

The task works by first partitioning reads into groups. Reads are grouped if they align to the same genomic position, have the same strandness, and any barcodes present match within an edit distance of two.

Within each group, sequenced UMIs are analyzed to determine whether they originated from the same UMI. To do this, UMIs are clustered. The UMI that has the most reads is used as the seed for the first cluster. The seed UMI is connected to all UMIs within a single edit distance that have fewer reads than it to form a cluster. Every UMI within the cluster then serves as the seed for a subsequent round of connection, again connecting seed UMIs to all UMIs within a single edit distance that have fewer reads than the seed UMI. Additional rounds of connection are performed until no more UMIs can be incorporated into the cluster. The unclustered UMI with the highest number of reads is chosen as the seed for a second cluster and the same clustering procedure is repeated. This process of clustering continues until all UMIs in the group have been assigned to a cluster.

This directional clustering method has two important benefits. First, it corrects for PCR errors introduced into UMIs by clustering sequenced UMIs with highly similar sequences so that they are counted as a single UMI. Second, it recognizes that PCR errors that arise in later cycles will be present in lower quantities. This is why clustering proceeds directionally, connecting UMIs with more reads to UMIs with fewer reads.

Once the clusters have been identified, a consensus read for the cluster is generated. To begin, any reads that do not match the common CIGAR string for their cluster are discarded. From the remaining reads, the percentage of each base at each position is determined. If a base is present in over 60% of reads, it is used in the consensus read. Otherwise, N is used. The base quality score for each position in the consensus read is the maximum at each position from the contributing reads.

Retain only one alignment per UMI

Deduplicate UMIs has an alternative setting to more closely match methods used by 10X Genomics' Cell Ranger pipeline: retain only one alignment per UMI. Selecting this option changes how the task functions and requires that you specify the genome assembly and gene/feature annotation.

The algorithm first checks whether each aligned read is compatible with a transcript in the annotation file. Here, compatible is defined as 50% or more of the aligned read sequence overlapping the transcript; strand is not considered. Aligned reads that are not compatible with a transcript or are mapped to multiple gene annotations are discarded.

The occurrence of each barcode and UMI combination is counted to establish the prevalence of each barcode+UMI.

UMIs within a Levenshtein distance of 1 are grouped.

The UMI within each UMI group with the highest number of reads is reported and other UMIs within the group are filtered out. If two UMIs within the group have the same number of reads, the UMI with the lowest sequence ASCII value is used.

This method is also similar to the default method in the Drop-seq Alignment Cookbook (Macosko et al. 2015), which collapses UMI barcodes with a Hamming distance of 1.

This method may output more UMIs than the default behavior as only UMIs within an edit distance of 1 are summarized, whereas UMIs with a greater distance can be linked in the UMI-tools method. For a comparison of the performance of the two approaches, please see the Adjacency (Cell Ranger) and Directional (UMI-tools) methods described in Smith et al. 2017.

References

Smith T, Hegar A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Research 2017; 27(3): 491-499. https://doi.org/10.1101/gr.209601.116

Connor, University of Michigan BRCF Bioionformatics Core https://github.com/umich-brcf-bioinf/Connor

Cell Ranger Algorithms Overview, 10X Genomics https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview

Drop-seq Alignment Cookbook v1.2 Jan 2016, James Nemesh, Steve McCarroll's lab, Harvard Medical School http://mccarrolllab.com/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf

Macosko E, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 2015; 161(5):1202-1214. https://doi.org/10.1016/j.cell.2015.05.002

Additional Assistance

If you need additional assistance, please visit our support page to submit a help ticket or find phone numbers for regional support.