QIAseq Ultraplex RNA Read Analysis

In this document, we describe the methods used for analyzing reads generated from the QIAGEN QIAseq Ultraplex RNA sequencing kit for single-cell or low-input RNA quantification. The analysis is broadly divided into primary and secondary analysis. Primary analysis generates UMI-based gene expression estimates from the reads. For single-cell applications, secondary analysis generates cell clustering and differential gene expression estimates.

Primary Analysis

Workflow

Step1: Demultiplex cells

The UMI and cell index sequences are extracted from the R2 reads. A specific read structure is expected from the R2 reads. The expected R2 read structure from a NextSeq instrument is:

<cell_index><UMI><ACG><poly-T>

because we use a custom sequencing primer that binds directly before the cell index. The expected R2 read structure from a MiSeq or HiSeq instrument is:

<AAGCAGTGGTATCAACGCAGAGTAC><cell_index><UMI><ACG><poly-T>

because we use the standard Illumina sequencing primer which binds before the QIAseq 25-mer uPCR adapter. At most 5 indel and mismatch errors in the uPCR adapter sequence and poly-T region are allowed. Only those reads which have a cell index with an exact match to one of the indices in our 96-cell or 384-cell index sets are retained.

If present, any 3' poly-A region is trimmed off the 3' end of the R1 reads. R1 reads with less than 25 base pairs are dropped.

Step 2: Align reads to reference genome

The poly-A trimmed R1 reads from the above step are aligned to the reference genome using the splice aware aligner STAR [1]. The reference genome used is GRCh38 (Human) or GCRCm38 (Mouse), to which ERCC spike-in sequences are added. Gencode Release 23 for Human and Gencode Release M15 for mouse was used to annotate splice junctions when building the genome index with STAR. The same transcript models are used for annotating reads during UMI counting.

Step 3: Count UMIs

The aligned reads from the above step are annotated to gene regions. UMIs are then counted for each gene across different cell indices. Before counting UMIs, we drop the following reads:

- For targeted enrichment, any R1 read with <= 25 bp aligned after the primer region (i.e. endogenous).
- For targeted enrichment, any R1 read which does not begin with a primer sequence.
- For targeted enrichment, any R1 read for which the primer is not aligned to the design locus.
- For poly-A transcriptome, any R1 read which is mapped to multiple genome loci.

Output Files

The results are provided in a zip file named "QIAseqUltraplexRNA_{job_id}.zip". The zip file contains (1) an Excel file called "QIAseqUltraplexRNA_{job_id}_run_summary.xlsx" which contains high-level summary metrics of both the primary and secondary analysis, a sub-folder containing primary analysis output files, and a sub-folder containing secondary analysis output files.

{job_id}.metrics.by_sample_index.txt: A tab delimited file containing metrics accumulated for each Illumina sample index. All the metrics below are mutually exclusive, i.e. the number of reads from the Illumina sequencer should be equal to the sum of all the reads used and all the reads dropped.

- **reads total:** Read pairs in the raw input fastq files.
- **reads dropped, all NNNNNN sequence:** Read pairs for which the R2 read contains only "N" base calls from the Illumina fastq generation.
- **reads dropped, cell id not extracted:** Read pairs dropped because the R2 read sequence did not follow the expected read structure (as described in the "workflow" section above).
- **reads dropped, cell id not matching oligo:** Read pairs dropped because the cell index extracted was not a perfect match to an index provided in the QIAseq cell index oligo plate set.
- **reads dropped, less than 25 bp:** R1 reads shorter than 25 bp.
- **reads dropped, not mapped to genome:** R1 reads not mapped to reference genome.
- **reads dropped, not annotated:** R1 reads not having any Gencode annotations based on the alignment location
- **reads dropped, off target:** R1 reads not aligned near a design region (targeted enrichment) or not aligned to a Gencode transcript model region (poly-A transcriptome).
- **reads dropped, primer not identified at read start**: R1 reads for which a primer sequence is not present at the read start position (targeted enrichment).
- **reads dropped, less than 25 bp endogenous sequence aligned after the primer:** Reads for which the read is not aligned to at least 25 bp immediately downstream of the designed primer binding site (targeted enrichment).
- **reads dropped, cell has no genes with more than 5 UMIs:** Reads corresponding to cells for which none of the genes has more than 5 UMIs.
- **reads dropped, aligned to genome, multiple loci:** Reads dropped because they are mapped to multiple genome loci (poly-A transcriptome).
- **reads used, aligned to genome, multiple loci:** Reads used for UMI counting which are mapped to multiple genomic regions (targeted enrichment).
- **reads used, aligned to genome, unique loci:** Reads used for UMI counting that are mapped to a single genome region.
- **reads dropped, aligned to ERCC, multiple loci:** Reads dropped because they are mapped to multiple ERCC spike-in sequences (poly-A transcriptome).
- **reads used, aligned to ERCC, multiple loci:** Reads used for UMI counting that are mapped to multiple ERCC-spike-ins (targeted enrichment).
- **reads used, aligned to ERCC, unique loci:** Reads used for UMI counting that are mapped to a single ERCC-spike-in sequence.
- **total UMIs:** Total number of UMIs in the gene count matrix.
- **mean reads per UMI:** Sum of all reads used divided by the total UMIs.

{job_id}.metrics.by_cell_index.txt: A tab delimited file containing metrics accumulated for each cell index.

- **cell identifier:** Cell identifier in the format <Illumina-index>_<QIAseq-cell-index>.

- **reads total:** Read pairs demultiplexed to each cell identifier.
- **reads used, aligned to genome:** Reads aligned to the reference genome used for UMI counting.
- **reads used, aligned to ERCC:** Reads aligned to ERCC-spike-ins used for UMI counting
- **UMIs:** Unique Molecular Index tags counted.
- **detected genes:** Number of genes with at least one UMI counted.

{job_id}.umi_counts.gene.{primer-pool-catalog-id}.txt: A tab delimited file containing UMI counts for each gene. For poly-A transcriptome read sets, all genes in Gencode Release 23 are listed. For targeted sequencing read sets, only genes targeted by the primer pool are listed. UMIs are summed across primers for a gene; i.e. a UMI is counted only once for a given gene when it has reads from more than one primer.

- **gene id:** Ensembl gene identifier.
- **gene:** HGNC gene symbol.
- **strand: '**1' when gene is on the "plus" reference genome strand ; '-1' when gene is on the"minus" strand.
- **chrom:** Chromosome of the gene.
- **loc 5':** Genome location of the 5' gene start (poly-A), or the 5' base of the primer binding site (targeted).
- **loc 3':** Genome location of the 3' gene end (poly-A), or the 3' base of the primer binding site (targeted).
- **Columns 7 onwards**: UMI count for each gene / cell index.

{job_id}.umi_counts.primer.{primer-pool-catalog-id}.txt: A tab delimited file containing UMIs for each primer in the primer pool (targeted sequencing). This file can be used for assessing primer performance, but should not be used for expression analysis because a sample fragment tagged with a UMI can be primed in the first PCR cycle with one primer, but then primed with a downstream primer (designed for a different transcript variant). So UMIs are sometimes double-counted in this file.

- **gene id:** Ensembl gene identifier.
- **gene:** HGNC gene symbol.
- **strand: '**1' when gene is on the "plus" reference genome strand ; '-1' when gene is on the"minus" strand.
- **chrom:** Chromosome of the gene.
- **loc 5':** Genome location of the 5' base of the primer binding site.
- **loc 3':** Genome location of the 3' base of the primer binding site.
- **primer seq:** Primer nucleotide sequence.
- **Columns 8 onwards:** UMIs corresponding to each primer / cell index.

Note: For low-input applications, an excel workbook with the above tables in separate sheets is provided instead of the individual text files.

Secondary Analysis

Workflow

Step 1: Quality control

The purpose of the QC step is to exclude "low quality" cells and genes that have extremely low read or UMI count. Specifically, the QC process takes two steps:

Drop cells with (1) number of reads mapped to endogenous genes below the 5th percentile, or (2) percent of endogenous gene reads below 5th percentile, or (3) number of detected genes below 5th percentile, or (4) fewer than 5 UMIs on from endogenous genes, or (5) fewer than 5 UMIs from all ERCC spike-in's (if available).

Drop genes with (1) fewer than 5 UMIs in total, or (2) fewer than 5 cells with at least one UMI, or (3) average of less than 5 UMIs per cell among cells with at least one UMI.

Step2: Normalization

The purpose of normalization is to remove technical variations such as different sequencing depths or RNA capture efficiencies. One of the following three normalization methods is used.

BASiCS [2] is the default method when ERCC spike-ins are available. The gene expression counts are normalized to size factors related to cell-specific technical bias and mRNA content. The parameters in BASiCS model are estimated with a Markov Chain Monte Carlo (MCMC) algorithm which may fail to converge in rare cases.

If BASiCS fails, gene expression counts are normalized by total ERCC spike-in UMIs in each cell, as implemented in "scran" [3].

If ERCC spike-ins are not available, gene expression counts are normalized by the total UMI count in each cell, as implemented by Zheng et al [4].

Step 3: Highly variable genes (HVG) identification

For many genes, the expression level varies from cell to cell, but this variation can be caused by real biological reasons and/or technical noise. In this step, we identify the genes of which a significant portion of cell-to-cell variation is attributed to biological reasons.

The default method is BASiCS. For each gene, the cell-to-cell expression variation is decomposed into three components: biological variation, technical variation, and sampling variation. HVGs are genes for which at least 40% of total variation is biological variation. The variance decomposition and HVG identification in BASiCS relies on successful convergence of the MCMC algorithm.

If the BASiCS MCMC algorithm fails, variance decomposition and HVG identification are performed using "scran". Using the assumption that most genes have roughly uniform expression level across the cells, a mean-variance trend is fitted on all endogenous genes and ERCC spike-ins and the fitted value of variance is considered as technical variation and the difference between total variance and fitted variance is considered as biological variation. HVGs are genes for which at least 40% of total variation is biological variation.

If ERCC spike-ins are not available, HVGs are not provided.

Step 4: Cell clustering

The purpose of cell clustering is to identify populations of cells. If HVGs are identified, cell clustering is based on the expression of HVGs. Otherwise, clustering is based on all genes. Log-transformation is performed on the normalized UMI counts to mitigate outlier effects. The clustering follows two steps:

Principal components analysis (PCA) is performed for dimension reduction. By default, we use the top 20 principal components.

The K-means algorithm is used for clustering. The default number of clusters (K) is predicted using the Gap statistic [5]. However, internal test data show that the prediction of K is not always correct, or consistent with visual impression from the first two dimensions in the PCA or T-SNE plots. Therefore, clustering results based on several possible K values are provided, ranging from K=2 to the minimum of 10 and 10% of the high quality cells.

The cells are visualized on a 2-dimensional T-SNE plot. T-SNE is a non-linear dimension-reduction method that maps the expression of hundreds of genes to a low dimensional space that can be more easily visualized. PCA, an alternative dimension reduction approach based on a linear transformation, is not as widely used as T-SNE in single-cell applications and is not provided in our analysis pipeline.

Step 5: Identifying marker genes

The purpose of marker gene identification is to find genes that are differentially expressed between two cell populations. SCDE [6] is used to perform differential expression between any two clusters and also between any cluster against all other cells.

Output Files

The output files are provided in a sub-folder named “secondary_analysis”.

{job_id}.step1_dropped_cells.csv: Low quality cells excluded from data analysis.

{job_id}.step1_dropped_genes.csv: Genes (including ERCC) excluded from data analysis due to low expression.

{job_id}.step1_RLE_ERCC.png: Relative log expression (RLE) values for all ERCC spike-ins in each cell. RLE is the ratio between the logarithm of a spike-in's UMI count in one cell and the logarithm of the median UMI count of that spike-in across all cells. Assuming all cells have the same ERCC input amount, the plot illustrates the variability of capture efficiency and RT efficiency across cells. This figure is generated only when there are less than or equal to 96 cells.

{job_id}.step2_normalized_UMI.csv: UMI counts for each gene / cell after normalization.

{job_id}.step3_highly_variable_genes.csv: Genes for which a large fraction (40% by default) of the total expression variability is explained by a biological cell-to-cell heterogeneity component.

{job_id}.step3_variance_decomposition.csv: Decomposition of total expression variability into three components:

- BioVarGlobal: biological cell-to-cell heterogeneity
- TechVarGlobal: technical variability
- ShotNoiseGlobal: baseline variance from Poisson sampling process

{job_id}.step3_CV_vs_mean_expression.png: Plot of log(CV) versus log(mean UMI) for each gene. The dashed line represents the baseline variance from Poisson distribution. Red markers are highly variable genes.

The sub-folders contain cell clustering and differential expression analysis results. Each sub-folder corresponds to a cluster number K. Within each sub-folder, the files are:

{job_id}.kmeans.clustering.csv: Cell clustering results using K-means method based on highly variable genes.

{job_id}.kmeans.clustering.png: Plot of first two dimensions of the T-SNE expression vector transformation. Cells are colored by their cluster membership according to the k-means clustering.

{job_id}.heatmap.png: Gene expression heatmap based on the normalized UMI counts after log transformation.

{job_id}.diff.exp.csv: Differential expression analysis results for each pair of clusters and for each cluster against other cells. All genes that are not dropped due to low expression are included in the analyses. In total, there are $\frac{K(K+1)}{2}$ such files and each file contains the following fields:

- K: number of clusters
- Group_A: first cluster in the pair
- Group_B: second cluster in the pair. In the one_vs_others analyses, it is "Others"
- Gene: gene name
- Fold_change: maximum likelihood estimate of fold change
- Lower_limit: lower bound of 95% confidence interval of the fold change estimate
- Upper_limit: upper bound of 95% confidence interval of the fold change estimate
- Z_score: un-adjusted z_score for testing significant fold change
- P_value: unc-adjusted p_value for testing significant fold change
- Adjusted_z_score: z_score adjusted for multiple testing
- Adjusted_p_value: p_value adjusted for multiple testing

The genes are sorted by Adjusted_p_value in ascending order.

References

[1] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras; STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, Volume 29, Issue 1, 1 January 2013, Pages 15–21, https://doi.org/10.1093/bioinformatics/bts635

[2] Vallejos, Catalina A., John C. Marioni, and Sylvia Richardson. "BASiCS: Bayesian analysis of single-cell sequencing data." *PLoS computational biology*11, no. 6 (2015): e1004333.

[3] Lun, Aaron TL, Davis J. McCarthy, and John C. Marioni. "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor." *F1000Research* 5 (2016).

[4] Zheng, Grace XY, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo et al. "Massively parallel digital transcriptional profiling of single cells." *Nature communications* 8 (2017): 14049.

[5] Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, no. 2 (2001): 411-423.

[6] Kharchenko, Peter V., Lev Silberstein, and David T. Scadden. "Bayesian approach to single-cell differential expression analysis." *Nature methods*11, no. 7 (2014): 740-742.