Quality Control Parameter

Standardized QC parameters were used to assess the quality of TaRGET II RNA-seq and ATAC-seq datasets. The following table contains detailed information on QC parameters used in TaRGET II RNA-seq and ATAC-seq pipelines.

Key QC parameters used in TaRGET II RNA-seq pipeline

|  |  |
| --- | --- |
| **QC Parameter** | **Explanation** |
| **uniq\_ratio** | The ratio between uniquely mapped reads and total input reads  |
| **assign\_reads** | The number of reads that were assigned to genes provided by GENCODE vM20 gene annotation file. |
| **genes\_w\_CPM\_gt1** | The number of expressed genes with CPM value larger than 1. |

Cutoff values for calculating QC score:

|  |  |  |
| --- | --- | --- |
| QC Parameter | cutoff value | QC score |
| >= cutoff | < cutoff |
| uniq\_ratio | 0.35 | 1 | 0 |
| assign\_reads | 10,000,000 | 1 | 0 |
| genes\_w\_CPM\_gt1 | 9,000 | 1 | 0 |

QC score used for determining data quality:

|  |  |  |
| --- | --- | --- |
|  | QC cutoff | data status |
| >= cutoff | < cutoff |
| sum of QC score | 3 | Pass | Fail |

Key QC parameters used in TaRGET II ATAC-seq pipeline

|  |  |
| --- | --- |
| **QC Parameter** | **Explanation** |
| **single\_end** | Useful single ends: the total number of useful single ends. This is each end of a non-redundant uniquely mapped read pair. |
| **enrp** | The promoters of active genes can be treated as positive control for open chromatin regions. ATAC-seq useful ends (E) at promoters overlapping ATAC-seq peaks are used to measure the signal enrichment:$$\frac{{E\_{under\\_promoter\\_peaks}}/{Length\_{total\\_promoter\\_peaks}}}{{E\_{total}}/{Length\_{genome}}}$$ |
| **enrs** | The ATAC-seq useful ends under ATAC-seq peaks are used to measure the signal enrichment at the genome-wide level. To avoid sequencing-depth bias, 10 million useful ends (E) are sampled from the complete dataset, and peak calling is performed to identify open chromatin regions. **enrs** is calculated after 10 million pseudo counts are added into the calculation as background, which can avoid calculation failure caused by the low sequencing depth:$$\frac{\frac{E\_{under\\_peaks}}{Length\_{total\\_peaks}}+\frac{10M}{Length\_{genome}}}{{(E\_{under\_{peaks}}+10M)}/{Length\_{genome}-Length\_{total\\_peaks}}}$$ |
| **rup** | The percentage of all useful ends that fall into the called peak regions with at least 50% overlap. $$\frac{number of useful ends under peaks}{number of total useful ends}$$ |
| **bk** | Fifty thousand genomic regions (500bp each) are randomly selected from the genome outside of ATAC-seq peaks. The ATAC-seq signal in each region is calculated as reads per kilobase per million mapped reads (RPKM). The percentage of all such regions with the ATAC-seq signal over the theoretical threshold (RPKM=0.377) is considered high-background and used as a QC metric to indicate the background noise:$$\frac{number of regions with high background}{50,000}$$ |

Cutoff values for calculating QC score:

|  |  |  |
| --- | --- | --- |
| QC parameter | value | QC score |
| single\_end | >=40000000 | 2 |
| [25000000, 40000000) | 1 |
| <40000000 | -1 |
| rup | >=0.2 | 2 |
| [0.12, 0.2) | 1 |
| <0.12 | -1 |
| enrp | >=11 | 2 |
| [7, 11) | 1 |
| <7 | -1 |
| enrs | >=18 | 2 |
| [15, 18) | 1 |
| <15 | -1 |
| bk | <=10 | 2 |
| (10, 20] | 1 |
| >20 | -1 |

QC score used for determining data quality:

|  |  |  |
| --- | --- | --- |
|  | QC cutoff | data status |
| >= cutoff | < cutoff |
| sum of QC score | 5 | Pass | Fail |