

From raw counts to differentially expressed genes

AN INTRODUCTION TO R PROGRAMMING & DESEQ₂
PACKAGE

RADIA JOHNSON, PH.D.

Outline

- Basics of R programming
 - Creating and working with vectors
 - Creating and working with matrices
 - Creating and working with factors
 - Creating and working with data frames
 - Creating and working with lists
- R script to analyse STAR gene count results
- Modify R script to analyse gene counts from HTSeq

quantmode

7 Counting number of reads per gene.

With `--quantMode GeneCounts` option STAR will count number reads per gene while mapping. A read is counted if it overlaps (1nt or more) one and only one gene. Both ends of the paired-end read are checked for overlaps. The counts coincide with those produced by htseq-count with default parameters. This option requires annotations (GTF or GFF with `-sjdbGTFfile` option) used at the genome generation step, or at the mapping step. STAR outputs read counts per gene into `ReadsPerGene.out.tab` file with 4 columns which correspond to different strandedness options:

column 1: gene ID

column 2: counts for unstranded RNA-seq

column 3: counts for the 1st read strand aligned with RNA (htseq-count option `-s yes`)

column 4: counts for the 2nd read strand aligned with RNA (htseq-count option `-s reverse`)

Select the output according to the strandedness of your data. Note, that if you have stranded data and choose one of the columns 3 or 4, the other column (4 or 3) will give you the count of antisense reads. With `--quantMode TranscriptomeSAM GeneCounts`, and get both the `Aligned.toTranscriptome.out.bam` and `ReadsPerGene.out.tab` outputs.

C:\Users\Radia\Downloads\hstseqUnion.sh - Notepad++

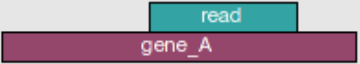
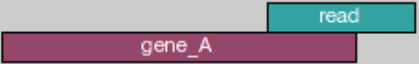


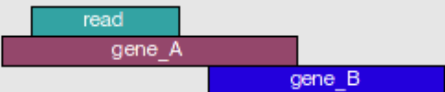

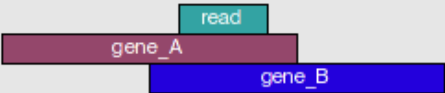
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?



microArrayEXMD.R x minimal_example.sh x star.sh x hstseqUnion.sh x

```
1  #!/bin/bash
2
3  # Add the modules needed for the analysis
4  module add HTSeq/0.6.1p1
5  module add SAMtools
6
7  # Variables
8  FILENAME=$1
9  myGTF="/RQexec/johnsonr/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf"
10 count=0
11
12 while read mySAMPLE
13 do
14     let count++
15     echo "$count $mySAMPLE"
16
17     # Go to the STAR directory in the PARK_LAB folder
18     cd $SCRATCH/STAR/"${mySAMPLE}"_STAR/
19
20     python -m HTSeq.scripts.count -m intersection-nonempty -f sam "${mySAMPLE}"Aligned.out.sam "${myGTF}" > "${mySAMPLE}.cnts"
21
22 done < $FILENAME
23
24 echo -e "\nTotal $count lines read"
25
26
27
```


HTSeq Options

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Options

- f** <format>, **--format**=<format>
Format of the input data. Possible values are `sam` (for text SAM files) and `bam` (for binary BAM files).
- r** <order>, **--order**=<order>
For paired-end data, the alignment have to be sorted either by read name or by read index. `name` indicates that the reads should be sorted by read name, and `index` indicates that they should be sorted by read index. The default is `name`.

If `name` is indicated, `htseq-count` expects all the alignments for the reads of a given name. If the mate of a read has not yet been seen, the read is kept in a buffer in memory until the mate is found. While, strictly speaking, this buffer is much less likely to overflow.
- s** <yes/no/reverse>, **--stranded**=<yes/no/reverse>
whether the data is from a strand-specific assay (default: `yes`)

For `stranded=no`, a read is considered overlapping with a feature regardless of whether it is on the same strand as the feature. For paired-end reads, the first read has to be on the same strand as the feature.
- a** <minqual>, **--a**=<minqual> ¶
skip all reads with alignment quality lower than the given minimum value (default: 10).
- t** <feature type>, **--type**=<feature type>
feature type (3rd column in GFF file) to be used, all features of other type are ignored.
- i** <id attribute>, **--idattr**=<id attribute>
GFF attribute to be used as feature ID. Several GFF lines with the same feature ID will be counted as one. For example, in an Ensembl GTF file, the attribute `gene_id` is used.
- m** <mode>, **--mode**=<mode>
Mode to handle reads overlapping more than one feature. Possible values for `<mode>` are `union`, `intersection_strict`, and `intersection_nonempty`.
- o** <samout>, **--samout**=<samout>
write out all SAM alignment records into an output SAM file called <samout>, and then exit.
- q**, **--quiet**
suppress progress report and warnings
- h**, **--help**
Show a usage summary and exit

Other resources

- R programming Blogs:
 - <https://rjbioinformatics.com/>
 - <https://www.r-bloggers.com/>
- R programming Books:

