mmquant

VERONIKA CHALUPOVÁ & HANA BOHÁČOVÁ

mmquant

- A tool to quantify gene expression

Published on: 15 September 2017 by Matthias Zytnicki

Last upload: 8 months and 28 days ago

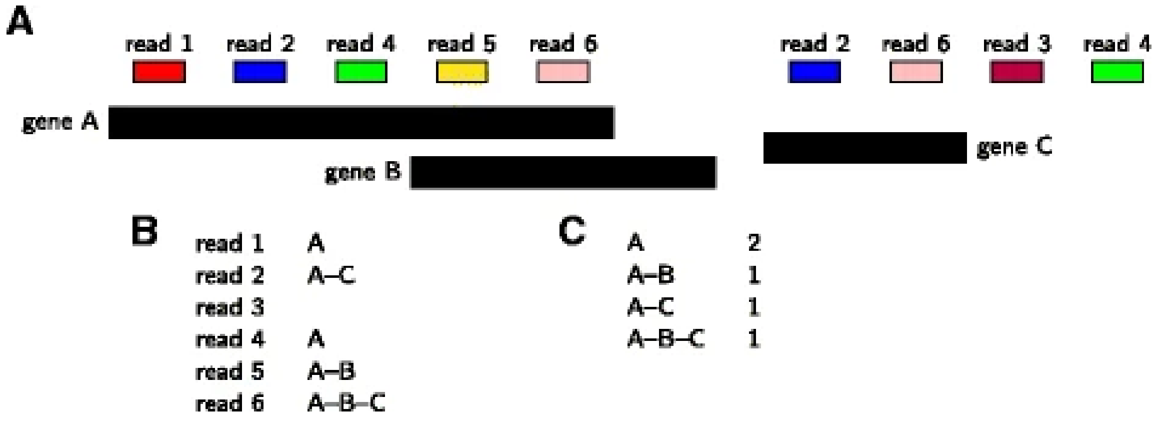
version: 1.3

Language used: C++

Operating system: Linux; Mac OS X

How does it work

- this tool counts with duplicated genes: if read maps to different positions, corresponding genes are duplicated -> this tool then creates a merged gene



read 1	A
read 2	A–C
read 3	
read 4	A
read 5	A–B
read 6	A–B–C

A	2
A–B	1
A–C	1
A–B–C	1

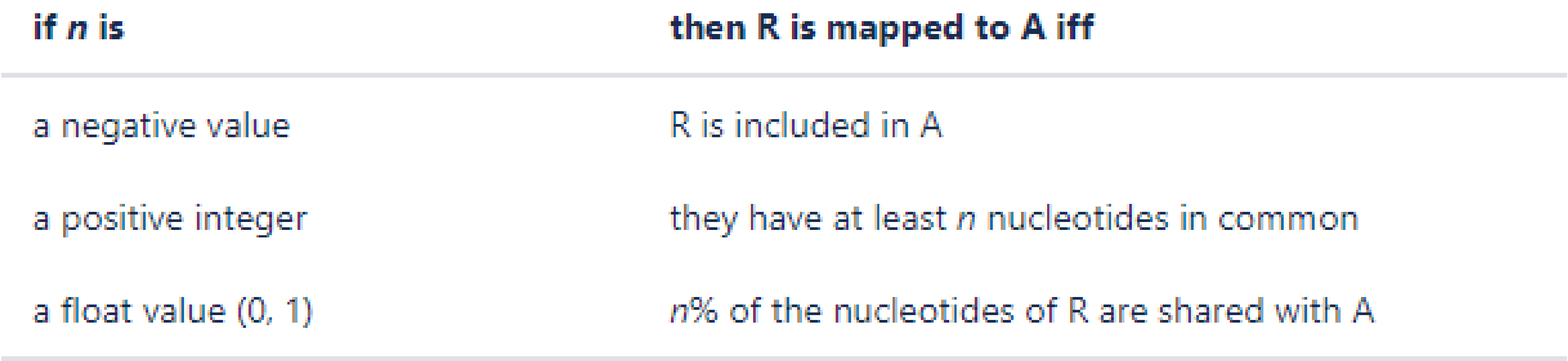
- by default, the method supposes that reads have been sorted beforehand

- if not: genes are sorted into a vector, cutted into non-overlapping bins and index is given to the first gene in bin; then for each read genes are scaned starting from first gene in bin

1. step of genes quantification

= searching for reads matching genes

The way a read R is mapped to a gene A depends on the -l *n* value set by user:



if *n* is	**then R is mapped to A iff**
a negative value	R is included in A
a positive integer	they have at least *n* nucleotides in common
a float value (0, 1)	*n*% of the nucleotides of R are shared with A

htseq-count: **union** / **intersection-strict** / *intersection-nonempty*
mmquant: **-l 1** / **-l -1** / *no alternative (ambiguous reads are discarded)*

- if read is mapped to several locations, the tool sets *NH* tag of SAM/BAM file to value >1

2. step of genes quantification

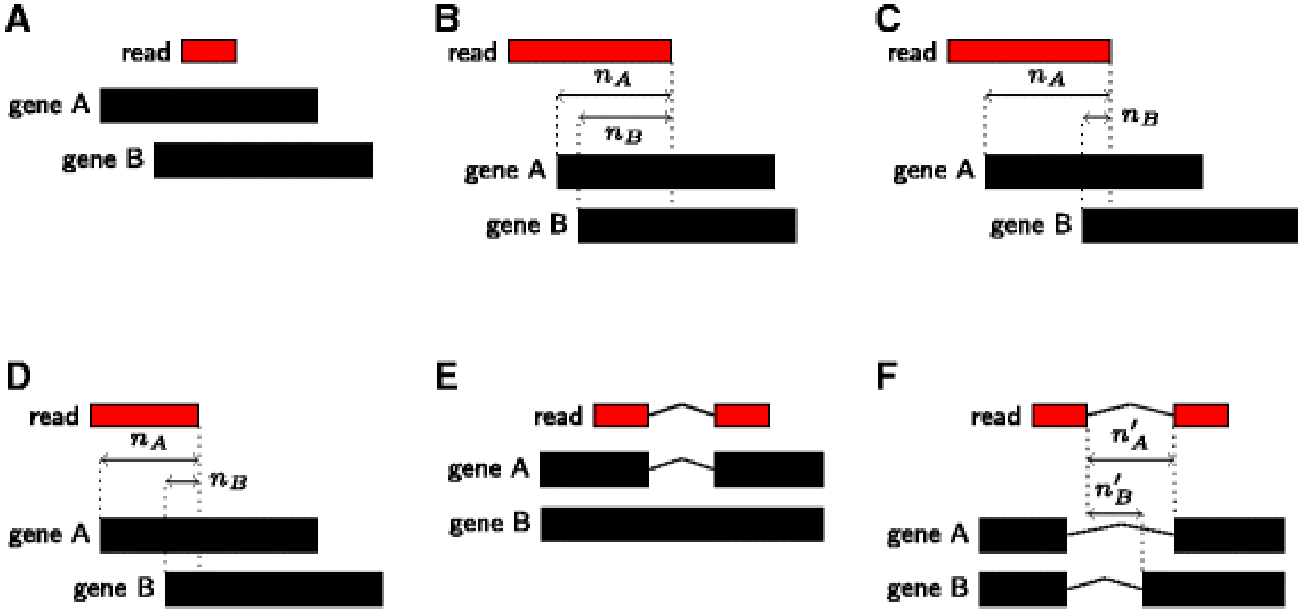
= resolving ambiguities

- when read matches several genes, some can be discarded depending on number of overlapping base pairs

-d *n* computes the differences of overlapping nucleotides (N_A, N_B). If N_A ≥ N_B + *n*, then the read will be attributed to gene A only.

-D *m* compares the ratio of overlapping nucleotides. If N_A / N_B ≥ *m*, then the read will be attributed to gene A only.

- featureCounts: option **largestOverlap** (assigns to the gene read with largest number of overlapping bases)
- mmquant: emulates this strategy by –d and –D parameters



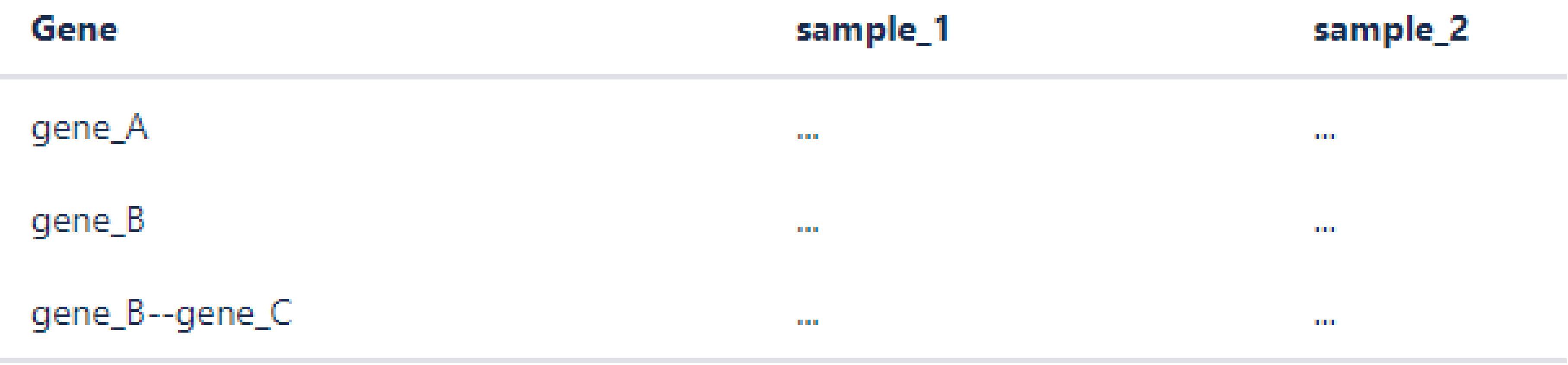
Input

Compulsory options:

- annotation file in GTF format
- reads in BAM/SAM format

Output

The output is a tab-separated file. It also provides output stats on hits.



Gene	sample_1	sample_2
gene_A	...	...
gene_B	...	...
gene_B--gene_C	...	...

Comparison with other tools

- time:
  - the fastest: featureCounts
  - also fast: mmquant
  - slowest: htseq-count

- number of expressed genes given by each tool is comparable
  - but multi-mapping genes could provide up to 25% of new genes
  - without them the results could be biased

Thank you for your attention