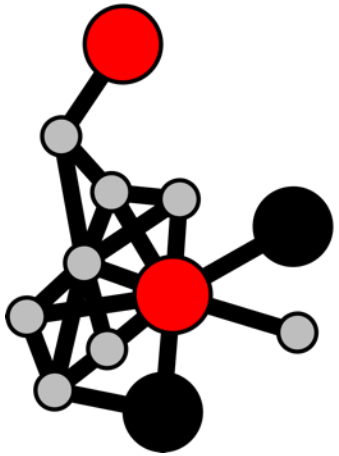
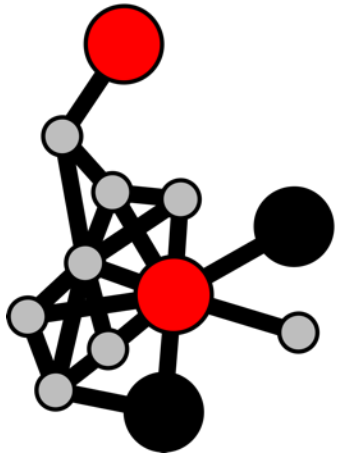


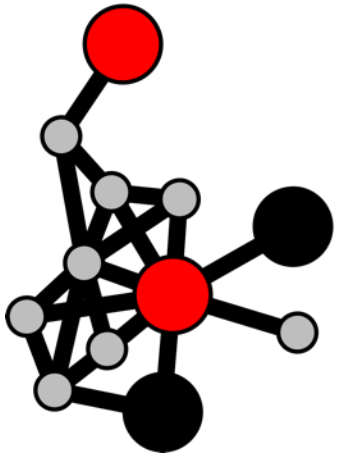


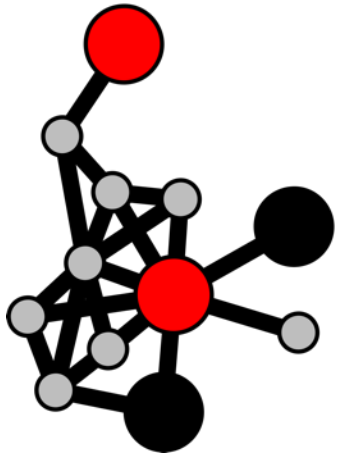
Transcriptomics lesson

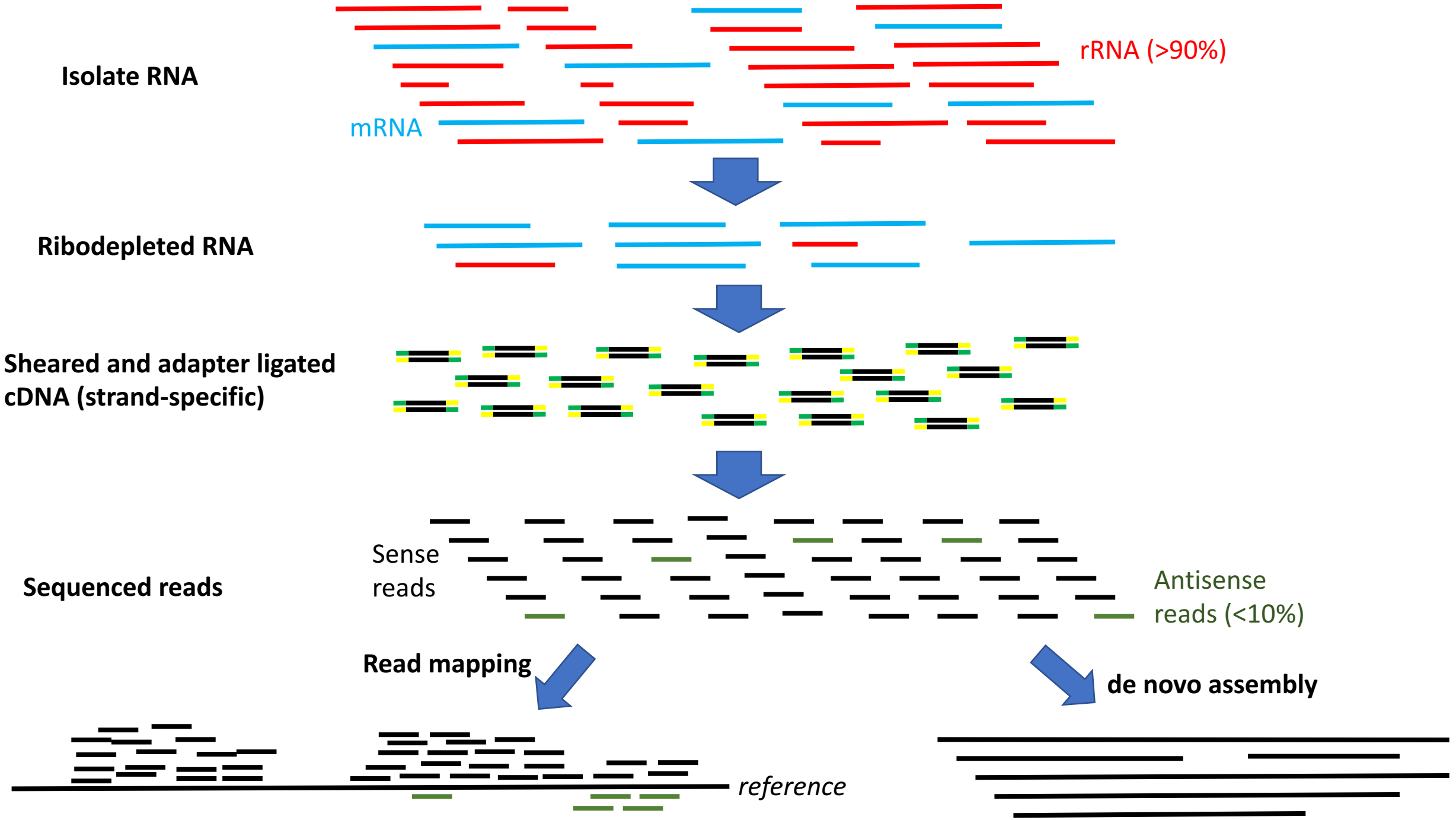
htseq-count:

Generating count tables from RNA seq read mapping

Launch binder






Isolate RNA
rRNA (>90%)
mRNA
Ribodepleted RNA
Sheared and adapter ligated cDNA (strand-specific)
Sequenced reads
Sense reads
Antisense reads (<10%)
Read mapping
de novo assembly
reference

OPEN ACCESS Freely available online

PLOS | ONE

RNA-Seq Gene Profiling - A Systematic Empirical Comparison

Nuno A. Fonseca*, John Marioni*, Alvis Brazma*

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, United Kingdom

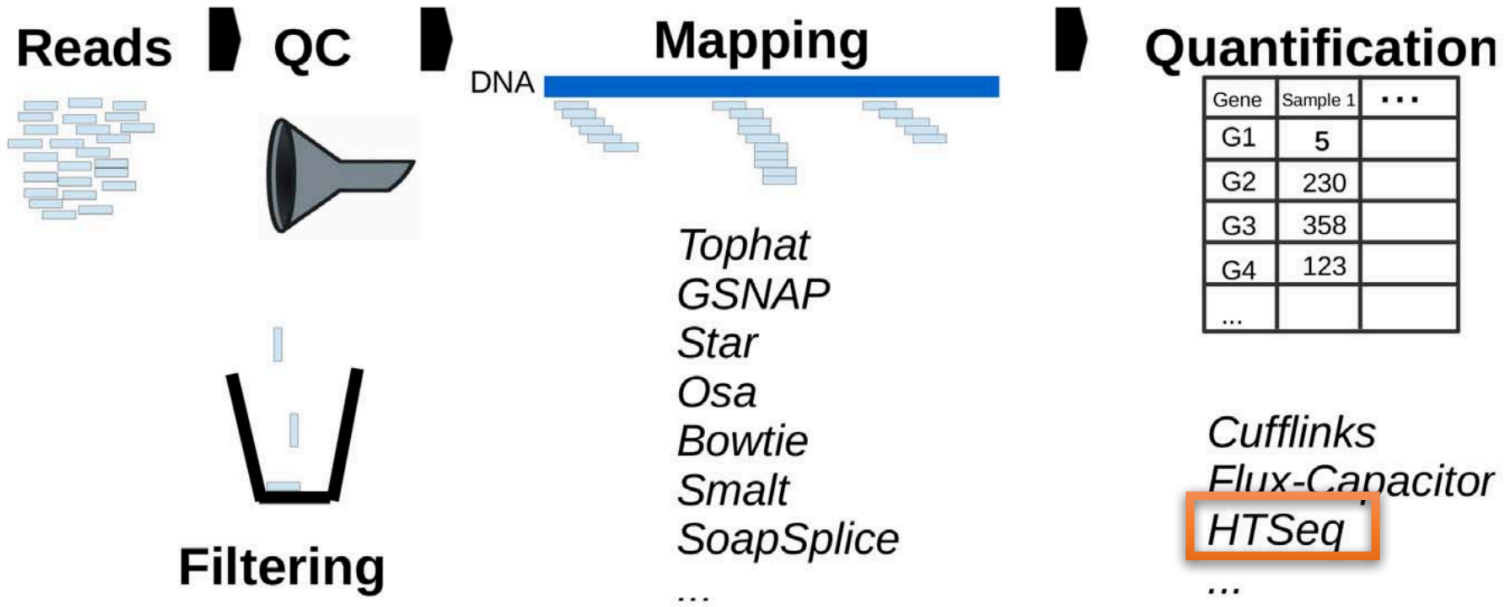


Figure 1. Gene profiling: from reads to gene expression.
doi:10.1371/journal.pone.0107026.g001

BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 2 2015, pages 166–169
doi:10.1093/bioinformatics/btu638

Genome analysis

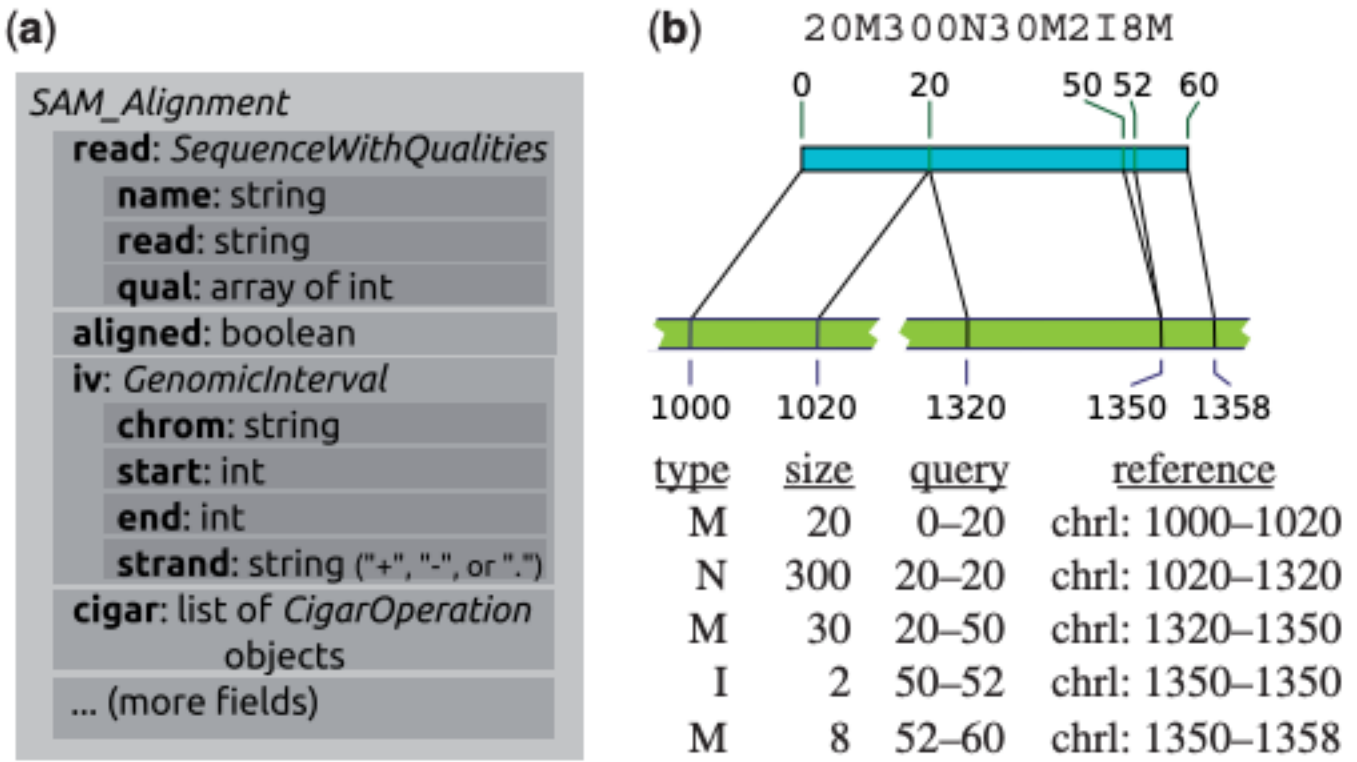
Advance Access publication September 25, 2014

HTSeq—a Python framework to work with high-throughput sequencing data

Simon Anders*, Paul Theodor Pyl and Wolfgang Huber

Genome Biology Unit, European Molecular Biology Laboratory, 69111 Heidelberg, Germany

Associate Editor: Michael Brudno



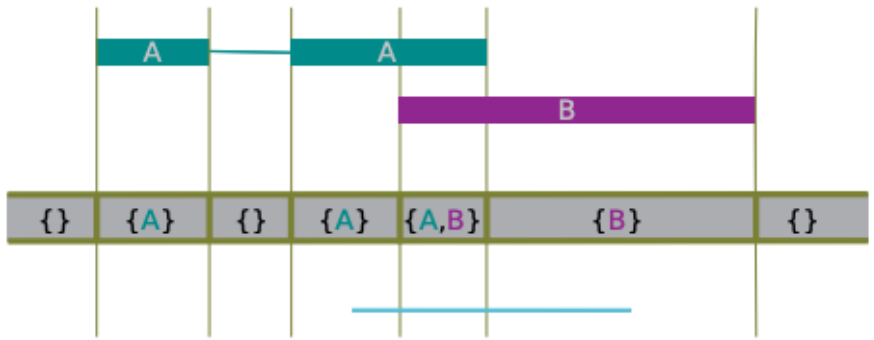
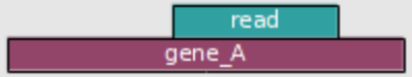
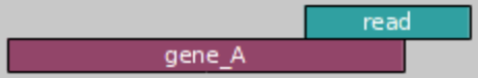
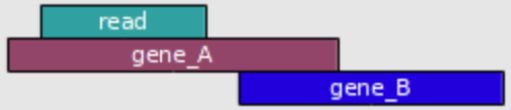


Fig. 2. Using the class *GenomicArrayOfSets* to represent overlapping annotation metadata. The indicated features are assigned to the array, which then represents them internally as steps, each step having as value a set whose elements are references to the features overlapping the step

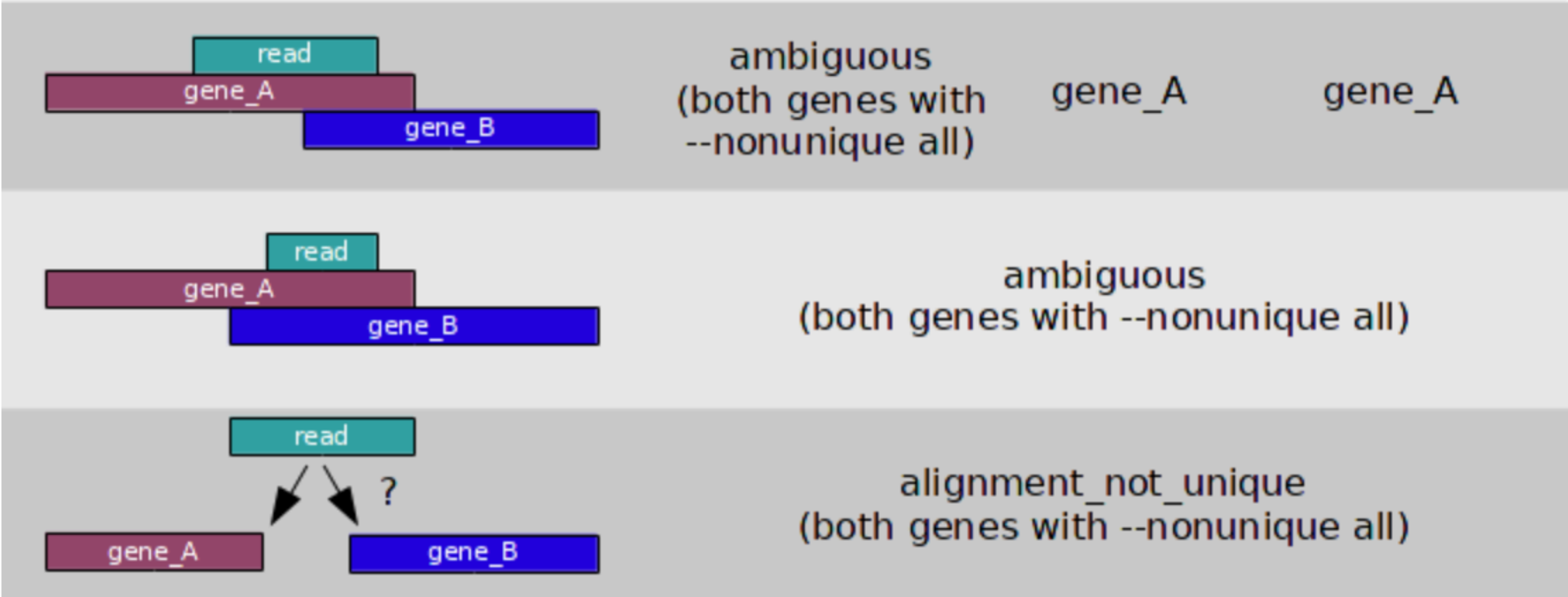
Read mapping parameters

	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

https://htseq.readthedocs.io/en/release_0.11.1/count.html

```
-m {union,intersection-strict,intersection-nonempty}, --mode {union,intersection-strict,intersection-nonempty}
                        mode to handle reads overlapping more than one feature
                        (choices: union, intersection-strict, intersection-
                        nonempty; default: union)
```

Read mapping parameters



https://htseq.readthedocs.io/en/release_0.11.1/count.html

```
--nonunique {none,all}
                        Whether to score reads that are not uniquely aligned
                        or ambiguously assigned to features
```

Read mapping parameters

https://www.ecseq.com/support/ngs/how-do-strand-specific-sequencing-protocols-work

NON-CODING RNA

Gene regulation by antisense transcription

Vicent Pelechano[1] and Lars M. Steinmetz[1–3]

Abstract | Antisense transcription, which was initially considered by many as transcriptional noise, is increasingly being recognized as an important regulator of gene expression. It is widespread among all kingdoms of life and has been shown to influence — either through the act of transcription or through the non-coding RNA that is produced — almost all stages of gene expression, from transcription and translation to RNA degradation. Antisense transcription can function as a fast evolving regulatory switch and a modular scaffold for protein complexes, and it can 'rewire' regulatory networks. The genomic arrangement of antisense RNAs opposite sense genes indicates that they might be part of self-regulatory circuits that allow genes to regulate their own expression.

~10% of gene expression is from antisense generally

```
-s {yes,no,reverse}, --stranded {yes,no,reverse}
                      whether the data is from a strand-specific assay.
                      Specify 'yes', 'no', or 'reverse' (default: yes).
                      'reverse' means 'yes' with reversed strand
                      interpretation
```

Counts tables designed for differential expression analysis (e.g. edgeR, DEseq2)

```
NC_009925.1_1   0
NC_009925.1_10  0
NC_009925.1_100 21
NC_009925.1_1000        35
NC_009925.1_1001        8
NC_009925.1_1002        2
NC_009925.1_1003        0
NC_009925.1_1004        2
NC_009925.1_1005        0
NC_009925.1_1006        9
NC_009925.1_1007        27
NC_009925.1_1008        68
NC_009925.1_1009        51
NC_009925.1_101 15
NC_009925.1_1010        13
NC_009925.1_1011        18
NC_009925.1_1012        23
NC_009925.1_1013        18
NC_009925.1_1014        77
NC_009925.1_1015        44
NC_009925.1_1016        54
NC_009925.1_1017        235
NC_009925.1_1018        29
NC_009925.1_1019        162
NC_009925.1_102 13
NC_009925.1_1020        26
NC_009925.1_1021        9
NC_009925.1_1022        93
NC_009925.1_1023        214
NC_009925.1_1024        31
NC_009925.1_1025        0
NC_009925.1_1026        0
NC_009925.1_1027        0
NC_009925.1_1028        0
NC_009925.1_1029        0
NC_009925.1_103 0
NC_009925.1_1030        1
NC_009925.1_1031        684
NC_009925.1_1032        1243
NC_009925.1_1033        29
NC_009925.1_1034        265
NC_009925.1_1035        1181
NC_009925.1_1036        568
NC_009925.1_1037        1016
NC_009925.1_1038        374
NC_009925.1_1039        5
```

Theory Biosci. (2012) 131:281–285
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 7 2014, pages 923–930
doi:10.1093/bioinformatics/btt656

Sequence analysis

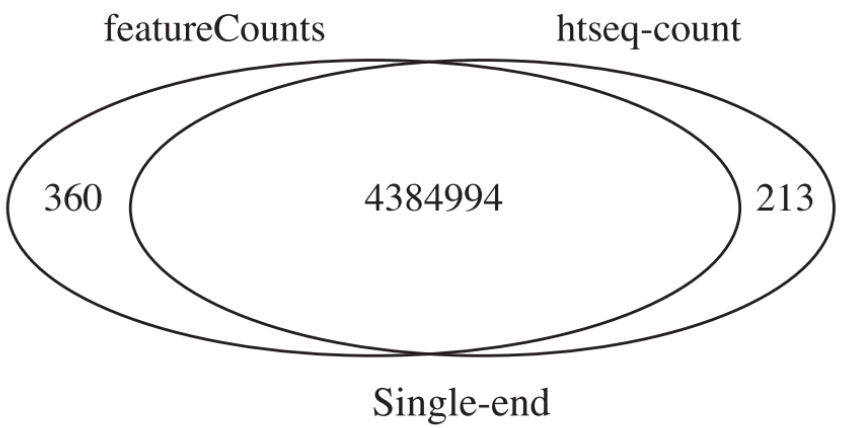
Advance Access publication November 13, 2013

featureCounts: an efficient general purpose program for assigning sequence reads to genomic features

Yang Liao[1,2], Gordon K. Smyth[1,3] and Wei Shi[1,2,*]

[1]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, [2]Department of Computing and Information Systems and [3]Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia

Associate Editor: Martin Bishop



Counting with `summarizeOverlaps`

Valerie Obenchain

Edited: August 2012; Compiled: August 23, 2013

BIOINFORMATICS ORIGINAL PAPER Vol. 30 no. 7 2014, pages 923–930
doi:10.1093/bioinformatics/btt656

Sequence analysis Advance Access publication November 13, 2013

featureCounts: an efficient general purpose program for assigning sequence reads to genomic features

Yang Liao[1,2], Gordon K. Smyth[1,3] and Wei Shi[1,2,*]

[1]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, [2]Department of Computing and Information Systems and [3]Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia

Associate Editor: Martin Bishop

Table 3. Performance with RNA-seq reads simulated from an annotated assembly of the Budgerigar genome

Methods	Number of reads	Time (mins)	Memory (MB)
featureCounts	7 924 065	0.6	15
summarizeOverlaps (whole genome at once)	7 924 065	12.6	2400
summarizeOverlaps (by scaffold)	7 924 065	53.3	262
htseq-count	7 912 439	12.1	78

Note: The annotation includes 16 204 genes located on 2850 scaffolds. *featureCounts* is fastest and uses least memory. Table gives the total number of reads counted, time taken and peak memory used. *htseq-count* was run in 'union' mode.

Table from Liao et al., 2014

Onto the Jupyter Binder Tutorial: htseq-count