Transcriptome Quantification

PRACTICAL

Gene/Transcript quantification - featureCounts

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

Liao Y, Smyth GK, Shi W.

Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656

Basic steps to analyse RNA-seq data

Mapping on reference genome

Transcript discovery
and/or
known transcripts

Quantification of gene/transcript expression

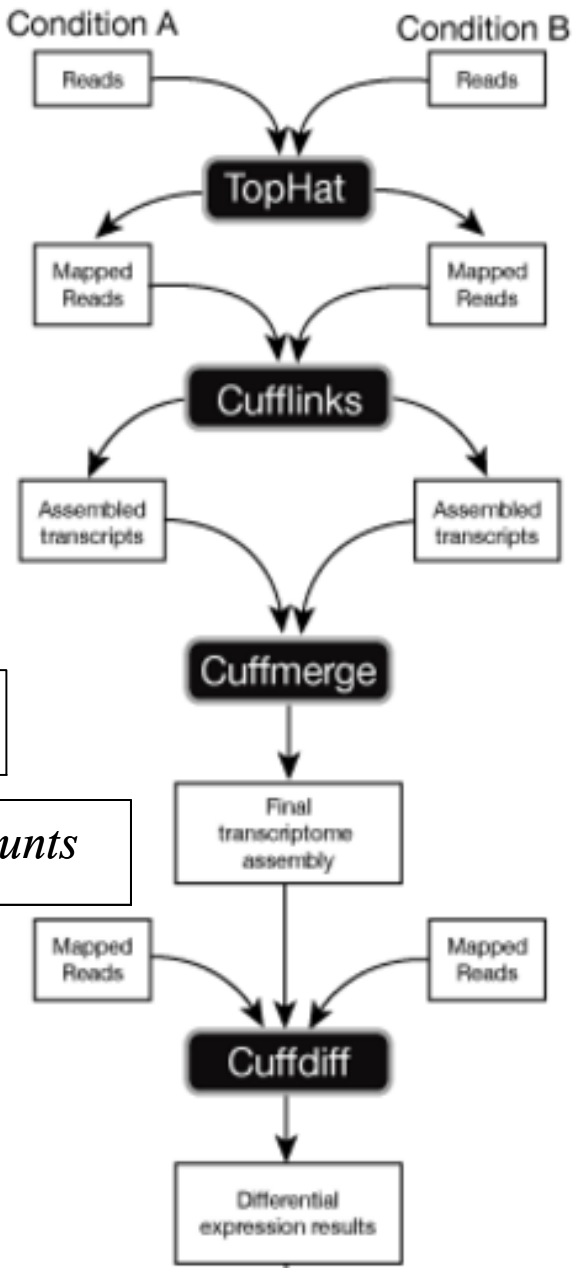
Differential analysis

featureCounts

raw counts

EdgeR,
DESeq,

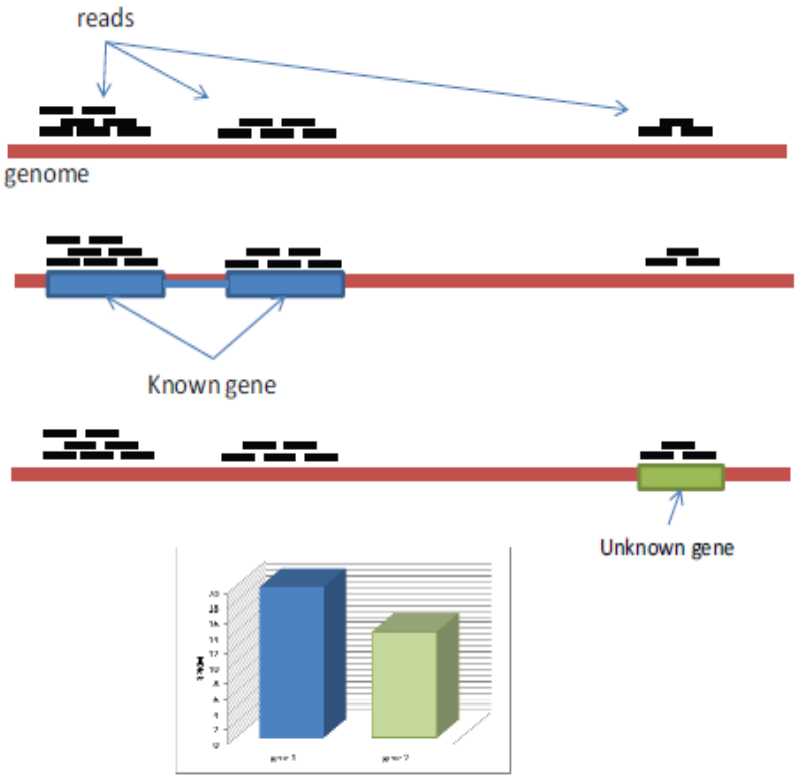
Tuxedo Suite



Gene/Transcript quantification - featureCounts

Gene/Transcript quantification is based on the number of reads (raw counts) mapping on it.

The simplest approach to quantification is to aggregate raw counts of mapped reads using programs such as HTSeq-count or **featureCounts**



Gene/Transcript quantification - featureCounts

- featureCounts is a program suitable for counting reads generated from either RNA or genomic DNA sequencing experiments

- faster than existing methods and requires far less computer memory

- single or paired-end reads supported (in the second case it counts fragments rather than reads)

- strand-specific read counting supported

- possible to specify a minimum mapping quality score that the assigned reads must satisfy
- highly flexible in counting multi-mapping and multi-overlapping reads, in fact such reads can be excluded, fully counted or fractionally counted

- it is part of the Subread (http://subread.sourceforge.net) or Rsubread (R package)

Gene/Transcript quantification - featureCounts

INPUT: one or more files of aligned read (SAM/BAM) and an annotation file (GTF/GFF/SAF) including chromosomal coordinates of features

OUTPUT: a count table, in which the number of reads assigned to each feature (or meta-feature) in each library is recorded

- a feature is an interval on one of the reference sequences:
  each entry in the provided annotation file is taken as a feature (e.g. an exon)

- a meta-feature is the aggregation of a set of features (e.g. a gene)
  the featureCounts program uses the gene_id attribute available in the GTF format annotation to group features into meta-features

→ one approach is counting reads overlapping each annotated exon in a GTF: can be used to test for alternative splicing between experimental conditions

→ another approach is counting reads at the gene level in a GTF: all reads that overlap any exon for each gene

PRACTICAL - featureCounts

```
cd $HOME/tutorial
mkdir featureCounts_2cells_output

#We want to count the reads mapping on each gene in the 2cells sample

/home/studente/Scrivania/Elixir-RNA-Seq-Tools/subread-1.5.3-Linux-x86_64/bin/featureCounts
-t exon -g gene_id -a /home/studente/Scrivania/Dataset_Corso/Danio_rerio.Zv9.66.gtf -o
featureCounts_2cells_output/counts.txt
/home/studente/tutorial/tophat_out/2cells/accepted_hits.bam

-t     specify the feature type: only rows which have the matched feature type (exon) in the
provided GTF annotation file will be included for read counting.
-g     specify the attribute type (gene_id) used to group features (e.g. exons) into meta-features
(e.g. genes).
-a     the name of the annotation file
-o     the name of the output file
```

#We want to rank the genes with respect to the number of mapped reads

```
awk '{print $1, $NF}' counts.txt | sort -gr -k 2 > 2cells_ordered_counts.txt

# awk prints only the gene id ($1: the first field of counts.txt file) and the counts ($NF: the last field)
# the output of awk is piped to be ordered (by sort command) with respect to the number of the reads (it is the
second field in the output of awk)
# the output of sort is written into a file
```

PRACTICAL - featureCounts

```
cd $HOME/tutorial
mkdir featureCounts_2samples_output

#We want to count the reads mapping on each gene in two samples

/home/studente/Scrivania/Elixir-RNA-Seq-Tools/subread-1.5.3-Linux-x86_64/bin/featureCounts -t
exon -g gene_id -a /home/studente/Scrivania/Dataset_Corso/Danio_rerio.Zv9.66.gtf -o
featureCounts_2samples_output/counts.txt
/home/studente/tutorial/tophat_out/2cells/accepted_hits.bam
/home/studente/tutorial/tophat_out/6h/accepted_hits.bam
```

-t specify the feature type: only rows which have the matched feature type (exon) in the provided GTF annotation file will be included for read counting.
-g specify the attribute type (gene_id) used to group features (e.g. exons) into meta-features (e.g. genes).
-a the name of the annotation file
-o the name of the output file