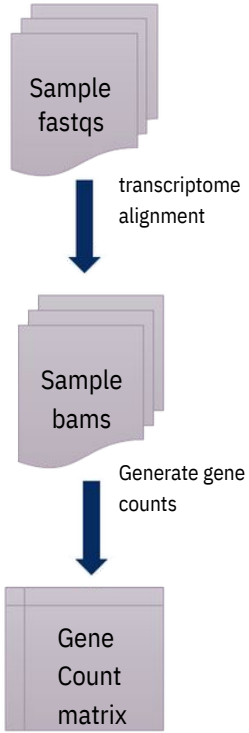
Preprocessing Transcriptomics data

Umar Ahmad, Ph.D.
[ORCID: 0000-0002-3216-5171]
Faculty Member, Sa'adu Zungur University

@babasaraky

Transcriptomics pre-proccesing workflow

Preprocessing



Extracting reads –different tools

- Subreads feature count
- Htseq RSEM
-

Summarize a BAM format dataset:

```
featureCounts -t exon -g gene_id -a annotation.gtf -o counts.txt mapping_results_SE.bam
```

Summarize multiple datasets at the same time:

```
featureCounts -t exon -g gene_id -a annotation.gtf -o counts.txt library1.bam library2.bam library3.bam
```

Perform strand-specific read counting (use '-s 2' if reversely stranded):

```
featureCounts -s 1 -t exon -g gene_id -a annotation.gtf -o counts.txt mapping_results_SE.bam
```

Summarize paired-end reads and count fragments (instead of reads):

```
featureCounts -p -t exon -g gene_id -a annotation.gtf -o counts.txt mapping_results_PE.bam
```

Summarize multiple paired-end datasets:

```
featureCounts -p -t exon -g gene_id -a annotation.gtf -o counts.txt library1.bam library2.bam library3.bam
```

Extracting reads –different tools

- Subreads_feature_count
- Htseq RSEM

```
htseq-count [options] <alignment_files> <gff_file>
```

-f <format>, **--format**=<format>
Format of the input data. Possible values are `sam` (for text SAM files) and `bam` (for binary BAM files). Default is `sam`.

-r <order>, **--order**=<order>
For paired-end data, the alignment have to be sorted either by read name or by alignment position. If your data is not sorted, use the `samtools sort` function of `samtools` to sort it. Use this option, with `name` or `pos` for <order> to indicate how the input data has been sorted. The default is `name`.

If `name` is indicated, `htseq-count` expects all the alignments for the reads of a given read pair to appear in adjacent records in the input data. For `pos`, this is not expected; rather, read alignments whose mate alignment have not yet been seen are kept in a buffer in memory until the mate is found. While, strictly speaking, the latter will also work with unsorted data, sorting ensures that most alignment mates appear close to each other in the data and hence the buffer is much less likely to overflow.

--max-reads-in-buffer=<number>
When <alignment_file> is paired end sorted by position, allow only so many reads to stay in memory until the mates are found (raising this number will use more memory). Has no effect for single end or paired end sorted by name. (default: `30000000`)

-s <yes/no/reverse>, **--stranded**=<yes/no/reverse>
whether the data is from a strand-specific assay (default: `yes`)

-m <mode>, **--mode**=<mode>
Mode to handle reads overlapping more than one feature. Possible values for *<mode>* are `union`, `intersection-strict` and `intersection-nonempty` (default: `union`)

Extracting reads –different tools

- Subreads feature count
- Htseq RSEM

```
htseq-count [options] <alignment_files> <gff_file>
```

-f <format>, **--format**=<format>
Format of the input data. Possible values are `sam` (for text SAM files) and `bam` (for binary BAM files). Default is `sam`.

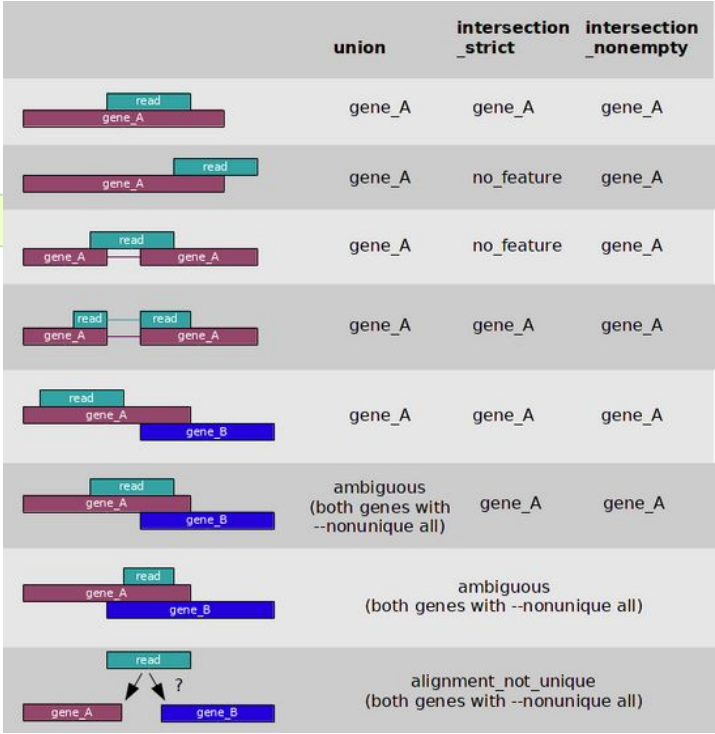
-r <order>, **--order**=<order>
For paired-end data, the alignment have to be sorted either by read name or by alignment position. If your data is not sorted, use the `samtools sort` function of `samtools` to sort it. Use this option, with `name` or `pos` for `<order>` to indicate how the input data has been sorted. The default is `name`.

If `name` is indicated, `htseq-count` expects all the alignments for the reads of a given read pair to appear in adjacent records in the input data. For `pos`, this is not expected; rather, read alignments whose mate alignment have not yet been seen are kept in a buffer in memory until the mate is found. While, strictly speaking, the latter will also work with unsorted data, sorting ensures that most alignment mates appear close to each other in the data and hence the buffer is much less likely to overflow.

--max-reads-in-buffer=<number>
When <alignment_file> is paired end sorted by position, allow only so many reads to stay in memory until the mates are found (raising this number will use more memory). Has no effect for single end or paired end sorted by name. (default: `30000000`)

-s <yes/no/reverse>, **--stranded**=<yes/no/reverse>
whether the data is from a strand-specific assay (default: `yes`)

-m <mode>, **--mode**=<mode>
Mode to handle reads overlapping more than one feature. Possible values for *<mode>* are `union`, `intersection-strict` and `intersection-nonempty` (default: `union`)



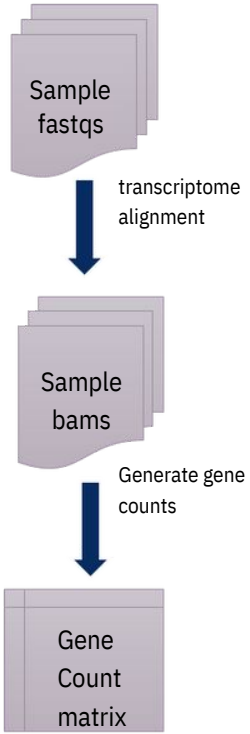
Extracting reads –different tools

- Subreads feature count
- Htseq RSEM

```
software/RSEM-1.2.25/rsem-calculate-expression -p 8 --paired-end \
                                               --bam \
                                               --estimate-rspd \
                                               --append-names \
                                               --output-genome-bam \
                                               exp/LPS_6h.bam \
                                               ref/mouse_ref exp/LPS_6h
```

Transcriptomics pipeline/workflow

Preprocessing



Transcriptomics pipeline/workflow

Preprocessing

Sample fastqs

transcriptome alignment

Sample bams

Generate gene counts

Gene Count matrix

Analyses

Clustering

Transcriptomics pipeline/workflow

Preprocessing

Sample fastqs

transcriptome alignment

Sample bams

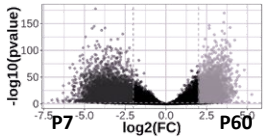
Generate gene counts

Gene Count matrix

Analyses

Clustering

Differential Expression



Transcriptomics pipeline/workflow

Preprocessing

Sample fastqs

transcriptome alignment

Sample bams

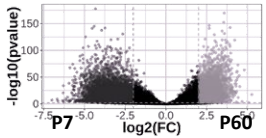
Generate gene counts

Gene Count matrix

Analyses

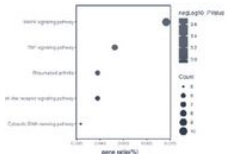
Clustering

Differential Expression



DESeq2 limma Glimma edgeR

Functional Enrichment



GSEA Gene Set Enrichment Analysis

MetaCore

GREAT

KEGG

reactome

DAVID BIOINFORMATICS DATABASE

Transcriptomics pipeline/workflow

Preprocessing

Sample fastqs

transcriptome alignment

Sample bams

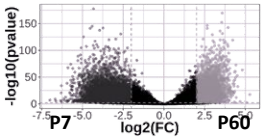
Generate gene counts

Gene Count matrix

Analyses

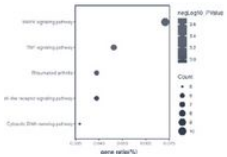
Clustering

Differential Expression



DESeq2 limma Glimma edgeR

Functional Enrichment



GSEA Gene Set Enrichment Analysis

MetaCore

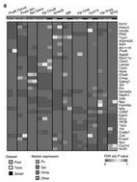
GREAT

KEGG

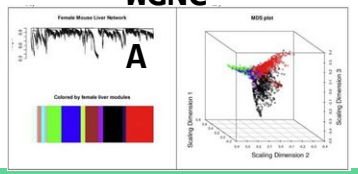
reactome

DAVID BIOINFORMATICS DATABASE

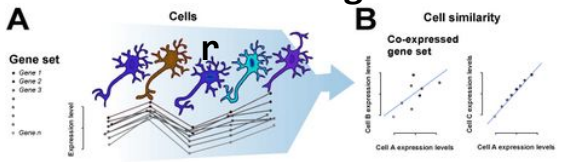
Coregulated Gene Expression



WGNC



MetaNeighbo



Questions?

Thank you for listening!