This protocol is based on performing differential expression analysis via the Galaxy server. Note: First and foremost, you should register at usegalaxy.org and activate your account.

Our pipeline consists of:

Raw FASTQ -> Trimmomatic -> HISAT2 -> FeatureCounts -> DESeq2

1. Upload your FASTQC files to the Galaxy server. Type and genome options are not necessary for this step, so leave them as default.
2. Search for “Trimmomatic” in the tools bar.
   1. Single-end or paired-end reads? –Paired-end (2 separate input files).
   2. Input FASTQ file (R1/first of pair)? – multiple datasets & select files ending with “-R1”.
   3. Input FASTQ file (R2/first of pair)? – multiple datasets & select files ending with “-R2”.
   4. Execute.
3. Download annotation files (GTF) and fasta files (.fa) from the ensemble bacteria database for desired strains and upload them to Galaxy.
4. Search for “HISAT2” in the tools bar.
   1. Source for the reference genome? –Use genome from history.
   2. Select the reference genome? –select the .fa file for desired samples.
   3. Is this a single or paired library? – Paired-end.
   4. FASTA/Q file #1? –multiple datasets & only R1 (paired).
   5. FASTA/Q file #2? –multiple datasets & only R2 (paired).
   6. Make sure that every pair is in order.
   7. Execute.
5. HISAT2 has a weird naming style, e.g., “HISAT2 on data x and data Y”, so feel free to change the name of the output in the history sidebar to something like “Control1”, “Media1” etc. If you do not know which data corresponds to which sample, you can click the HISAT output, and it will show you the pairs that were used to produce this output.
6. Search for “featurecounts” in the tools bar.
   1. Alignment file? –multiple datasets & select all BAM files from HISAT2.
   2. Gene annotation file? –in your history & choose the appropriate GTF file you uploaded earlier in step 3.
   3. Execute.
7. Search for “DESeq2” in the tools bar.
   1. Specify a factor name, e.g., effects\_drug\_x or cancer\_markers? –up to you
   2. Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'? –up to you.
   3. Counts files? -multiple datasets & only “Counts” files (not summary) for 1st group of samples (for example 3 treated counts files).
   4. Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'? –up to you.
   5. Counts files? -multiple datasets & only “Counts” files (not summary) for 2nd group of samples (for example 3 control counts files).
   6. Go to output options and select all available types of outputs.
   7. Execute.
8. DESeq2 plots file is downloadable and contains all the graphs for your differential gene expression analysis.
9. DESeq2 results file is a huge table with all the statistical data that you can use to determine the significance of the individual results and whether the gene is upregulated or downregulated.
10. But the first search for “Annotate DESeq2/DEXSeq” in the tools bar.
    1. Tabular output of DESeq2/edgeR/limma/DEXSeq? – select a single DESeq2 “results” file of your choice
    2. Reference annotation in GFF/GTF format? –select the appropriate GTF file you have uploaded earlier in step 3.
    3. Execute.
11. Step 10 will produce an “Annotate DESeq2” results table that contains all the statistical data + at what chromosome gene is located (column 8), start and end positions of the gene (column 9 & 10), which strand (column 11), gene function (column 12) and gene name (column 13).
12. Download the file mentioned in step 11 and change the extension to xlsx, as it is easier to manipulate.
13. Add headers to columns in the following order:
    1. GeneID
    2. Basemean
    3. Log2FC
    4. Stderror
    5. Wald
    6. P-value
    7. Adjusted p-value
    8. Chromosome location
    9. Start
    10. End
    11. Strand
    12. Function
    13. geneName
14. In excel go to data>filter and:
    1. Filter adjusted p-value by deselecting NA from the list.
    2. Sort genes by log2FC (does not matter whether it is ascending or descending).
15. Upregulated genes have log2FC > 0, whilst downregulated genes have log2FC < 0.
16. You can further filter this file by setting adjusted by value to less than or equal to 0.05. The resulting table will contain only genes that experience significant differences.