## Brief for the Skill Clip

|  |  |
| --- | --- |
| **Goal Video:** | Which part of the protocol and what will it explain. |
| The general goal of the video is to educate a wider audience about coding in R and R studio to tackle big data and quickly analyze it to gain significant insights in, for example, sample characteristics. We performed DESEQ2 analysis to analyze lung cancer data and implemented gene enrichment analysis to link the significant differentially expressed genes (DEGs) to the specified disease. Through this video, we strive to show the pipeline implemented in our analysis and what is possible with bioinformatics. | |
| **Audience:** | Who is the audience of this video? |
| The target audience for this video is the people who are interested in doing bioinformatics and computational analysis of big data sets. It is meant as an introductory level video to help people with either a general beginner level of understanding or no level of understanding at all to navigate through the majestic world of bioinformatics. | |
| **Topic:** | What is the video topic? Be specific. |
| The video created is about the pipeline of bioinformatic lung cancer analysis. The video introduces the disease of non-small cell lung cancer (NSCLC), explaining why it is important to analyze. Next, it shows how to retrieve the lung cancer data from the GEO database, and by installing certain libraries/packages through R, analysis can start. First,t it shows the exploratory analysis to, as a goal, find out more about the data e.g. its expression levels. Next, it shows how to load the count matrix, which allows us also to load the metadata which is essentially the sample information. Whenever these are uploaded, the video shows how to create a DESeq2 dataset. This allows us to perform quality control and get the DESeq2 results, sorted by p-value and log2foldchange. The p-value has to be adjusted to account for the large dataset and thus allows us to retrieve and effectively filter for the significant genes, shown by the video. Lastly, the video shows the realized plots in the form of volcano plots which show up- and downregulated genes & use MOET for functional enrichment analysis. This allows us to link the DEGs to known diseases. | |
| **Key takeaways:** | What are the key takeaways of the video? What should viewers learn? |
| The viewers should be educated on the basics of bioinformatics and how to use the coding language R, with its interface R studio. By implementing these skills, viewers should get a general idea of how to tackle the analysis of big data files to insights into e.g. sample characteristics or DEGs in specified disease profiles. | |
| **Call-to-action:** | What is the call-to-action? (in this case to replicate the protocol) |
| Whenever you want to replicate the protocol, do make sure to follow the steps outlined to ensure proper reproducibility for the analysis. Please ensure that all dependencies/libraries are installed, that metadata is correctly formatted, and that the sample selection criteria are applied as described. Run the provided functions step by step to perform the DESeq2 analysis successfully. Make sure to use GitHub to track your steps whenever errors occur & to enhance collaboration with colleagues. | |

## Script

Sound: any spoken text, sounds, or music please add them here.

Image: this can be simply the presenter talking or performing a skill, a test set-up, a slide, a schematic representation, a text bullet, or an image that clarifies your story.

Remarks: any extra remark

Time: This does not need to be very exact. It is good however to time your story. Please aim to do a 2 to 4 minute clip, no longer than 5 minutes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sound**  **(Describe any sound, music, narration, dialogue, sound effects, or any sound)** | **Image -shot & action**  **(Scene- Describe anything visible)** | **Remarks**  **(any extra remarks for the camera team or editor)** | **Time**  **approx** |
| Note: Music used throughout the video is: <https://www.youtube.com/watch?v=9KOmsZkukM0>. It has no copyright. | Note: The presenter is in the shot in every scene, unless not specified | Note for editor: Generally, move around the presenter in such a way that the slide content is always visible. | Note: The first time refers to the time of the single scene. The one in brackets is the total added time (all scenes up there). |
| Music starts from this slide on. | Introduction slide with our logo and names with no presenter. | A nice transition would be appreciated. | 0’04’’  (0’04’’) |
| *Hello everybody, we are group 18, and welcome to our skills video.* | Introduction slide with our logo and names. The presenter along with background music fades in. |  | 0’04”  (0’08”) |
| *Lung cancer is one of the deadliest cancers in the world. The most prevalent type of lung cancer is non-small cell lung cancer (NSCLC).*  *By understanding its molecular mechanisms it is possible to identify potential therapeutic targets to tackle this horrendous disease.* | Lung cancer (LC) diagram illustrating the two main types of LC.  Focus on NSCLC in the LC diagram, emphasizing NSCLC by circling it in the diagram. |  | 0’16’’  (0’24”) |
| *In this video, we will learn how to analyze RNA-seq data.*  *Our* ***goal is*** *to identify differentially expressed genes (DEGs) between lung cancer and adjacent healthy tissues. To do so we will use programming language R, a versatile toolbox designed for statistical analysis.* | Slide in which we write our objective.  The presenter is in the shot. | Create an overlap with the background slide and Rstudio slide. | 0’09”  (0’33’’) |
| *Within the R toolbox, we can find specialized toolsets, each containing tools (functions) and materials (datasets) tailored for specific tasks which we call “packages”.* | Diagram in which the R logo contains a toolset. From the toolset, two arrows come out, one that points to a building tool and the other to a building material (e.g. wood). |  | 0’09”  (0’42”) |
| *The informatic pipeline we will follow today specifically makes use of the package DESeq2.*  *DESeq2 is a software package that allows us to identify differentially expressed genes from RNA-Seq data.*  *it follows 4 steps: For each gene, it models the raw counts to account for differences in library depth. This is also referred to as the normalization step. It then estimates the gene-wide dispersions. Finally, it fits these data into a statistical model, which is the negative binomial model. This allows it to perform hypothesis testing using the Wald test. The output will be a p-value used to determine the significant DEGs ([DEsEq source](https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html)).*  *If you are interested in the topic we highly suggest checking the linked video (*[*https://www.youtube.com/watch?v=0b24mpzM\_5M*](https://www.youtube.com/watch?v=0b24mpzM_5M) *) But now, let’s see how we practically perform differential expression analysis.* | Show the informatic pipeline of the whole project. Then show the diagram of DESeq2. | Make sure in the slide there is a link to the video referred to in the script. | 0’48”  (1’30’’) |
| *For our study, we use publicly available RNA-seq data from the Gene Expression Omnibus (GEO) database, specifically this dataset \*. It contains gene expression profiles of tumor tissue from “199 patients and 19 paired normal lung tissues”.* | Show the GEO webpage, in which the dataset G*SE81089 is open.*  \* The presenter points at the dataset \* | In the GEO page, highlight the line referring to the “199 patients and 19 paired paired normal lung tissues” part. | 0’19”  (1’49”) |
| *From the GEO webpage, we can directly download raw counts tables or the normalized FPKM counts. Both data files are essentially counts matrices in which rows represent genes and columns are samples. The numerical values represent the number of reads mapped to that gene in that sample.* | A slide showing the raw counts file. |  | 0’19”  (2’08”) |
| *For the FPKM counts, the value is normalized based on gene length and the total number of mapped reads* | A slide showing the FPKM file. |  | 0’05”  (2’13”) |
| *Additionally, we can also find the metadata file, containing any additional characteristics of patients.*  *Let’s now proceed with the analysis of this data.* | A slide showing the metadata file. |  | 0’10”  (2’23”) |
| *We will follow 9 steps of which the first 5 comprise preparing and performing DESeq2 analysis, whilst the last ones are needed for visualization and interpretation of data* | Subsequently, it shows a quick recap of the slide that demonstrates the general DESeq2 pipeline. |  | 0’09”  (2’32”) |
| *Our requirements include our dataset, and then R and RStudio, which is the integrated development environment for R.* | Show FPKM data and the R studio logo. |  | 0’10”  (2’42”) |
| *We will also need these R packages. So when we opened Rstudio to install these packages, we used BiocManager, which is a package contains all the packages mentioned. Before proceeding any further, we load them all.*  *Once our libraries are loaded, we load FPKM and raw data counts of our dataset into R.* | Show all packages used and their general functions. | At the end, show the piece of code in which we install and load all required libraries. | 0’18”  (3’00”) |
| *Before moving on, we still need to perform some exploratory analysis to check our data 2) identify possible issues*  *and 3) gain initial insights*  *To do so we used the already normalized FPKM data.* | Slide about exploratory analysis in which the goals of these steps are elucidated. | Cut the parts in which the presenter reads all the questions. | 0’14’’  (3’14’’) |
| *We plot a Manhattan plot to see the distribution of gene expression across the genome. In this way, we check if selected cancer genes are indeed expressed and how they are approximately distributed across the genome.* Once we had a quick look at our data and everything seems to work, we can proceed to create a DESeq2 dataset. | Show the Manhattan plots. |  | 0’18”  (3’32”) |
| *We use the function DESeqDataSetFromMatrix. Here, we need to discuss three important points.* | Show the function of how to create a DESeq2 dataset object. | Line 148 in our code | 0’08’’  (3’40”) |
| *Firstly, we are required to give an input datafile (countData). Which will be the raw data and not the normalised one. If you recall the DESeq2 workflow, in fact, normalisation is automatically performed in the first step.* | Show function in which the first line is highlighted. | Refer somehow back to the DESeq2 workflow please | 0’15”  (3’55”) |
| *Secondly, in the next line, we are integrating the raw counts data with the sample metadata. Importantly, to do so, we need to make sure that the columns of the raw data match with the rows of the metadata file* | Show slide in which the second line of the function is highlighted.  Here, the alignment requirement (other snippets of code in which we align columns and rows) should be present. |  | 0’10’’’  (4’05”) |
| *Lastly, the design formula sets the factor of interest to test for.*  *So summarising, we are constructing a DESeqDataSet object (dds) that integrates the raw counts from the meta data with the sample metadata. This will be analysed according to the variable “source” which is nothing else but the health status* | Last line of the function code is emphasised. Here it is visually explained (written down part of the script) what this is used for. |  | 0’25’’  (4’30”) |
| *We performed quality control by removing low expression genes by keeping only those with an average count of 10 or more across all samples. This helps with the accuracy of the analysis.* | Slide with line of quality control and then after |  | 0’10’’  (4’40”) |
| *By running deseq, we get the log2 change, which is calculated for in the design factor in which the healthy tissue is set as control. So positive values mean overexpressed compared to healthy. Looking at the other columns of the matrix, the base mean is the average of the normalized counts taken over all the samples. The lfcSE column is the standard error for the log2 change. And the pvalue and adjpvalue are calculated for the stat value (Wald test) for the gene indicated on the rows.* | Output of DESeq2 results. | If you can emphasise the variables discussed. | 0’30”  (5’10”) |
| *Sorting by pvalue, we keep the significant DEGs which are thousands. Therefore we adjust the threshold to 0.01.* | Slide with lines of code that filter out significant DEGs. |  | 0’05”  (5’15”) |
| *We can visualize the results using a volcano plot. In this way we can clearly see that some genes are up - and downregulated. Higher points indicate more significant genes due to the lower p-value.* | Slide with the Volcano plot with downregulated and upregulated genes. |  | 0’15’’  (5’30”) |
| *Finally, we perform, functional Enrichment Analysis to identify biological processes, molecular functions, and cellular components that are overrepresented in our DEGs set.* | First a slide with an explanation of functional enrichment analysis. |  | 0’10”  (5’40”) |
| *We used Multi Ontology Enrichment Analysis (MOET). We found 5596 DEGs. With this method, we can narrow down the amount of DEGs and focus on the most biologically relevant genes involved in lung cancer.* | A slide of the MOET output when adding 500 DEGs. |  | 0’15’’  (5’55”) |
| *In summary, we used RNA-seq data from lung cancer patients to identify differentially expressed genes and analyze their involvement in key biological pathways. By leveraging bioinformatics tools such as R, DESeq2, and MOET, we developed a robust pipeline for gene expression analysis. These findings contribute to a better understanding of lung cancer and may support future biomarker discovery and therapeutic advancements.* | Final slide in which the concluding diagram is shown. | Here, it might be useful to refer back to the previous slides (GEO, MOET and DESeq) | 0’30”  (6’25”) |
| *Thank you for watching! We hope this video has given you a clear understanding of RNA-seq analysis and bioinformatics workflows.* |  | Add a concluding transition please. | 0’05’’  (6’30’’) |