RNAseq reads to differential genes and pathways

27 & 28 September 2022

Archive of questions from the workshop Slack Channel

(In the order they were asked during the workshop)

Questions about the pipeline and RNAseq

Question/comment	Answer
Can you explain allele specific expression more?	This is about differences in expression between alleles - e.g. expression of a gene inherited from one parent compared to the other. This analysis pipeline is different to the one we will go through today
Are these column headings for the sample sheet are must-have?	Yes, there is a naming convention that the programs used for RNAseq data follow.
So is it convention that fastq_1 always the forward?	Yes. They can also be L1 and R1
If you have your own reference genome and annotate it using say example Augustus, then how do you use it in the nextflow?	If you have your own reference genome and annotate it using say example Augustus, then how do you use it in the nextflow. in the nextflow command, you can change the parameters to use your own reference and annotation files with the --fasta and --gtf options
Does Nextflow have pipelines for RNA-seq analysis of prokaryotes?	you can browse nf-core pipelines, which can be found here: https://nf-co.re/ There are currently 69 pipelines available here
Is there a list of reference genomes that are available onNextflow?	It depends on the specific workflow you're using. If you go to https://nf-co.re/ you can see all the available pipelines and read more about each pipeline Also, nf-core give you access to Illumina's

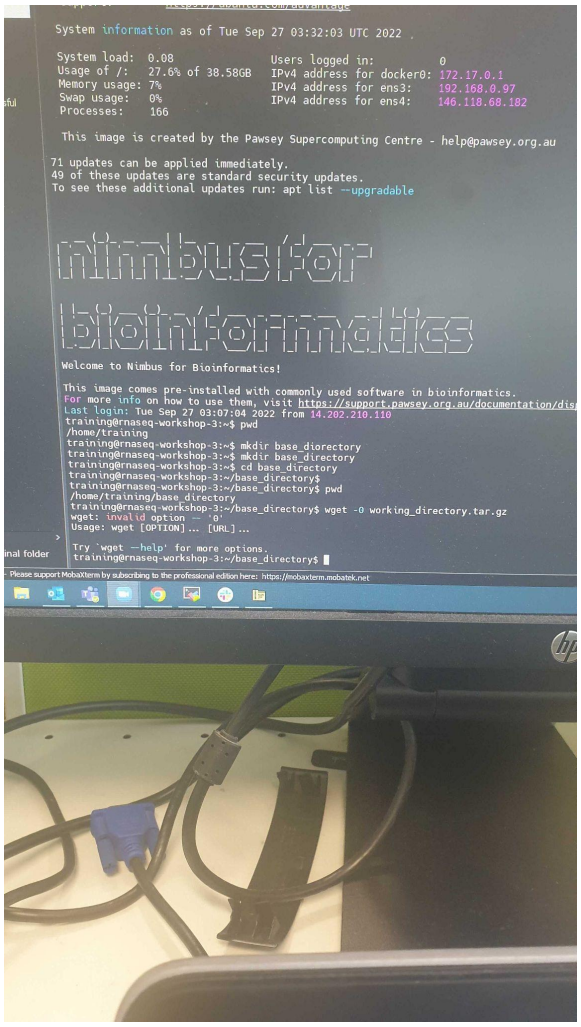
	iGenomes database https://nf-co.re/usage/reference_genomes which has pre-stored and indexed reference genomes for a few species.
in regards to the `-profile singularity` option, is singularity a tool we also need to install?	Singularity is a software that runs containers. Nextflow handles the installation for you
My model organism is a cow and i dont see a cow reference genome in the reference genome database. So all i have to do is download the genome sequence and the annotation files from RefSeq in the working directory?	Yes, that's right.
In terms of customizing the workflow, if I like to adjust certain flags for some of the steps, for example for STAR, adjusting the read length taken into the pipeline or adjust the strand specific flags, how can I adjust specific parameters please?	The link to nf-co.re/rnaseq parameters: https://nf-co.re/rnaseq/3.8.1/parameters
Day 2	
Can we do parallelization (for example alignment of all samples at the same time) inside nextflow?	Yes, if you have sufficient compute resources nextflow can run many jobs at once We did that yesterday. Download the timeline.html report from your run yesterday to see which processes were able to run in parallel. Each process run by nf-core has specific cpu and memory requirements and your VM has a set number of CPUs and memory. Nextflow will run as many processes as it can in parallel given the resources available to it on the machine.
Can we use normalized/standardized gene expression data for DE analysis?	Depends on the tool you use. Today we will use DESeq2 which requires raw counts. This is because DESeq2 does normalization internally (and has it's own unique method to do so).
In the nextflow of nf-core, what is the difference between the fasta and Star-index file? My understanding is star-index are used for STAR alignment and quantification, while fasta is used for Salmon to generate	Index files are used by tools (like STAR, salmon) to perform look-up functions faster (i.e. when it is aligning to a reference genome, it can do this faster). The index method/files are tool specific

salmon index file for counts and possible for bam etc.? Is that right?	Sure, but the index file are normally come from the reference genome, right? I guess from metagenomics, we use the assembly for both index and fasta, maybe different in human genome, as the index can be very specific sets of genes only. That's right - the index file is specific to the reference genome and tool (and even tool version). You should generate index files with your FASTA reference file & tools that you use
Is there a list of R studio containers that are ready to be used through nimbus please?	RStudio is installed already on Nimbus if you set up your instance using the bio image. See this documentation for more info on how to open it: https://support.pawsey.org.au/documentation/display/US/Nimbus+for+Bioinformatics#NimbusforBioinformatics-RStudio See also: https://support.pawsey.org.au/documentation/display/US/Run+RStudio+Interactively We're working with a container we developed specifically for today's training, you can access it at the same cvmfs path if you'd like to use it for yourself and we plan to release it publicly soon.
Why do we use Rstudio to do DE rather than a nextflow pipeline?	DE analysis (rather than data processing as we did on day 1) involves a lot more interactive analysis, looking at plots, etc. So, it is a nicer way to perform this analysis. I would add, yesterday, we were processing our data from raw sequence reads to generate a dataset that could be analysed. Its ok to automate processing steps which are generally more standardised across experiments compared with analysis that changes a lot depending on your research questions.
Why does DESeq2 uses a negative binomial distribution model?	Take a look at their documentation, go down to Theory behind DESeq2 in http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

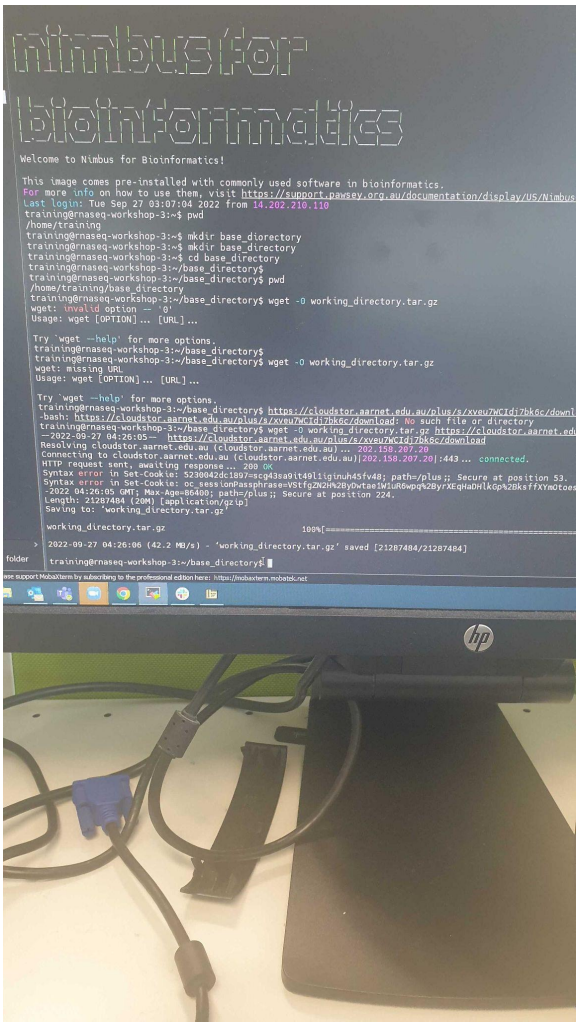
	Here is an example of a weird dispersion plot : https://support.bioconductor.org/p/107937/ And a discussion as to why. Michael Love is one of the DESeq2 authors and is actively on the bioconductor forum if you want to ask the authors a question. Otherwise do ask a local statistician/bioinformatician
Isn't p.adjust the same as qvalue?	Yep, the qvalue package doing slightly different clusterprofiler is just offering you a bit of choice in the types of padj methods you can apply, whereas the qvalue applied here is just the false discovery rate
Why do we conduct DE for up- and down-regulatory genes separately?	It's much easier to interpret this way. We can identify which group of genes/enrichments are upregulated together in the knockout as compared to the wild, and vice versa

Error messages

Question/comment	**Answer**
When connecting to Nimbus my connection timed out	Try connecting to a different internet network (e.g. Eduroam). Some institutional internet connections block access to Nimbus.



Ah ok, its meant to be -O (the letter O) not the number 0

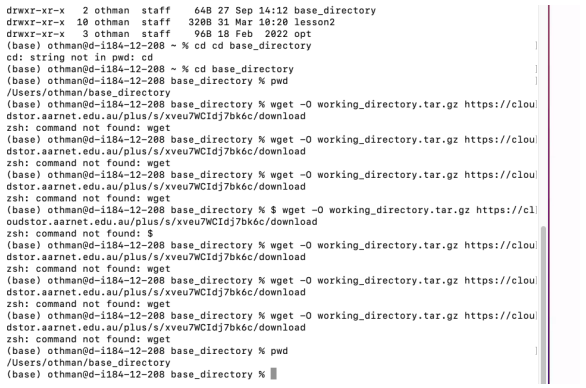


Looks like the file downloaded successfully. You can check by typing the command `ls`

After `ls` do `tar -zxvf working_directory.tar.gz`

then `cd working_directory`

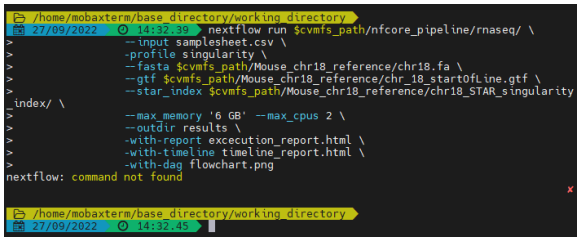
those commands 1. unzip the file you just downloaded 2. change your directory to working_directory



You aren't logged into your nimbus vm. it appears you're on your local computer.

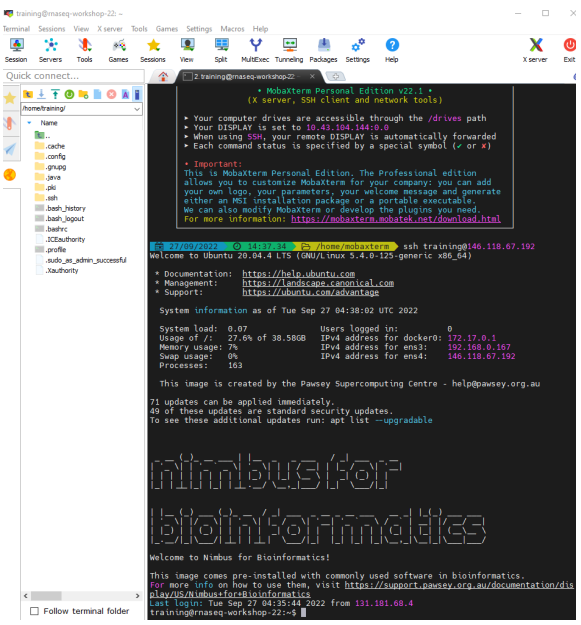
Follow directions here to login https://sydney-informatics-hub.github.io/rna-seq-pt1-quarto/setup.html#connect-to-nimbus

I get command not found



Looks like you are not logged onto Nimbus.

follow directions here to login https://sydney-informatics-hub.github.io/rna-seq-pt1-quarto/setup.html#connect-to-nimbus



You are logged in now, just make sure to keep working in this terminal window.

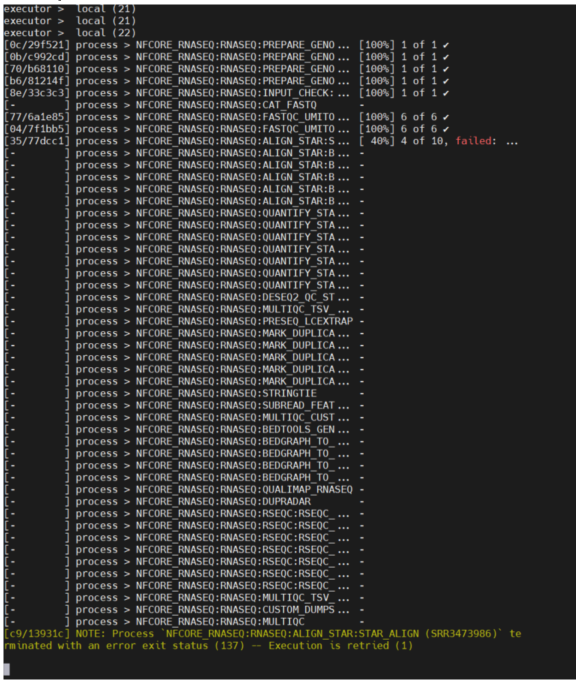
my command failed: [bd/239935] NOTE: Process `NFCORE_RNASEQ:RNASEQ:ALIGN_STAR:STAR_ALIGN (SRR3473985)` terminated with an error exit status (137) -- Execution is retried (1)

you can try to rerun your pipeline by adding -resume to the end of the command. Like this:

```
nextflow run
$cvmfs_path/nfcore_pipeline/rnaseq/ \
                --input samplesheet.csv
\
                -profile singularity \
                --fasta
```

```
$cvmfs_path/Mouse_chr18_reference/chr18.fa \
                --gtf
$cvmfs_path/Mouse_chr18_reference/chr_18_startOfLine.gtf \
                --star_index
$cvmfs_path/Mouse_chr18_reference/chr18_STAR_singularity_index/ \
                --max_memory '6 GB'
--max_cpus 2 \
                --outdir results \
                -with-report
excecution_report.html \
                -with-timeline
timeline_report.html \
                -with-dag flowchart.png
-resume
```

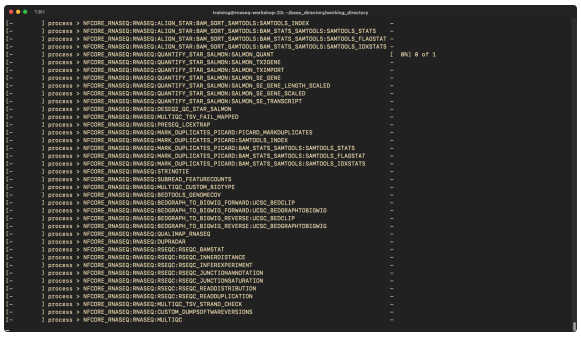
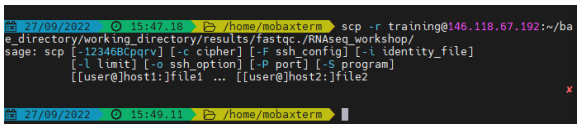
My command keeps failing at a specific fastq

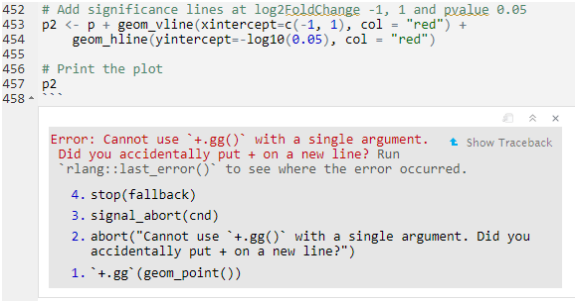


pipeline completed with warning

nf-core/rnaseq] Pipeline completed successfully-
WARN: To render the execution DAG in the required format it is required to install Graphviz -- See http://www.graphviz.org for more info.

that's to be expected! Nandan will talk about that when we go through the output

It seems like the command from the previous step didn't run for me. I'm stuck with this screen 	try `ctrl + c` and then try to rerun the command
`training@rnaseq-workshop-45:~/base_directory/working_directory $ cd results` `training@rnaseq-workshop-45:~/base_directory/working_directory/results$ ls -lh results` `ls: cannot access 'results': No such file or directory` `training@rnaseq-workshop-45:~/base_directory/working_directory/results$ cd results` `-bash: cd: results: No such file or directory` `training@rnaseq-workshop-45:~/base_directory/working_directory/results$ ^C` `training@rnaseq-workshop-45:~/base_directory/working_directory/results$`	you're already in the results directory
how do i get ls-lh to work?	it's `ls -lh` with a space in the middle
Has anyone have permission issue to transfer data? 	you've only got train@ you need to have training@ I would recommend only downloading the fastqc and trimgalore folders, There are some big alignment files in there
	You are missing a space just before ./RNASeq_workshop/ there is no "." in fastqc

Day 2	
I don't have the warning about the package on my RStudio. Is this something to worry about?	
I'm not getting a new "ready" line after running the password command	that means it's working! now go to chrome browser then Type 146.118.XX.XX:8787 in your browser where the XX.XX will be replaced by your IP specific digits:
I don't get the warning when I run the 'DESeqDataSetFromMatrix' chunk, will assume this is fine but unsure why it's different from Nandan's	Neither do I, but I can run all the code fine so we can probably disregard it.
Just getting an error on the volcano plot  ``` # Add significance lines at log2FoldChange -1, 1 and pvalue 0.05 p2 <- p + geom_vline(xintercept=c(-1, 1), col = "red") + geom_hline(yintercept=-log10(0.05), col = "red") # Print the plot p2 ``` Error: Cannot use `+.gg()` with a single argument. Did you accidentally put + on a new line? Run `rlang::last_error()` to see where the error occurred. Show Traceback 4. stop(fallback) 3. signal_abort(cnd) 2. abort("Cannot use `+.gg()` with a single argument. Did you accidentally put + on a new line?") 1. `+.gg`(geom_point())	I think there's a typo somewhere, can you compare/use this code: ``` # Create a basic volcano plot (scatter plot) with x axis = LogFC, y axis = -log10(pvalue) resdata <- as.data.frame(res) # Define whether genes are significantly DE or not and store this in a new column called DE resdata$Significant <- "No" resdata$Significant[resdata$log2FoldChange > 1 & resdata$pvalue < 0.05] <- "Upregulated" resdata$Significant[resdata$log2FoldChange < -1 & resdata$pvalue < 0.05] <- "Downregulated" # Create the volcano plot p <- ggplot(data=resdata, aes(x=log2FoldChange, y=-log10(pvalue), col=Significant)) + geom_point() # Add significance lines at log2FoldChange -1, 1 and pvalue 0.05 p2 <- p + geom_vline(xintercept=c(-1, 1), col = "red") + geom_hline(yintercept=-log10(0.05), col = "red") # Print the plot p2 ```
My knitting error is as follows: `No protocol specified`	looks like you have a View() command in one of your code blocks.

Quitting from lines 568-603 (rnaseq_DE_analysis_Day2.Rmd) Error in .External2(C_dataviewer, x, title) : unable to start data viewer Calls: <Anonymous> ... withCallingHandlers -> withVisible -> eval -> eval -> View In addition: Warning messages: 1: In DESeqDataSet(se, design = design, ignoreRank) : some variables in design formula are characters, converting to factors 2: Removed 8 rows containing missing values (geom_point). 3: In View(sig.up) : unable to open display Execution halted	Try hash it out or remove it and then knit again.