

Analysis of bulk RNA-seq data

18th to 20th November 2020

9:30am - 17:30pm (GMT) (Hopefully not quite that late!)

Links

All the materials for the course, as well as the full timetable can be found at the [Course Website](#)

- **Please connect to the course via this Zoom link:**
<https://us02web.zoom.us/j/89596597815?pwd=TFMveVMyaUxwTjhodDEyWHVrRFZwZz09>
- **Course participant introductions:**
<https://docs.google.com/document/d/1Fuydp4llf7fuz6Gn816EES9vH3eWc4w3eVJec>
- **Course feedback survey:** <https://www.surveymonkey.co.uk/r/8DLF62C>
We would really appreciate it if you could share your thoughts with us regarding these sessions. We are interested in your opinions, how you feel the experience has benefited you and how it could be improved.
If you could find a few minutes to complete a short survey at the end of the last session it would really help us in improving the training we can deliver.
- **Mentimeter :** <https://www.mentimeter.com/>
- **Participants introduction:** [e9kGyE/edit?usp=sharing](https://www.mentimeter.com/e9kGyE/edit?usp=sharing)
- **Some of the tools used:**
 - fastQC: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Further actions:

- [Sign up to our mailing list to get notifications of upcoming courses from the Bioinformatics Training Facility](#)

Recordings

Recordings for each day's teaching will be posted here shortly after the zoom call closes. These recordings are intended for course attendees only for private study only. Please do not share these links. The links will continue to work for around a month after the course but you may download the videos for a permanent copy if you wish.

- Day 1

<https://us02web.zoom.us/rec/share/fOlyHd68iEQ3u3VHugxeXlgmWhzeNs2yGaQVzd-LQQxB1mB5o2ixfewR471pB-Q.wjPP96H3H4SKZga2> Passcode: i!uuvK87

- Day 2

https://us02web.zoom.us/rec/share/zm21gOJm_VT3LoqEtFc4Dlp0lyk9r2HNuOC1u_m0hzgJzPwBtNoJggs8mazKhrxDm.bzw99u9NoyEGuaNq Passcode: CU?\$pFe2

- Day 3

https://us02web.zoom.us/rec/share/UgXjilBHh9vwsH5ohK_xFhuom-7YWc-aUP3StxbIh_AibYIVO2f35anhiZy-F6Um.7m3Lnue7tQ9jlQw9 Passcode: D56#T6+&

Remote Machines

During the course we will use two remote servers to run both command line tools and R. Shortly before the start of the course you should receive an email with links, a username and a password.

Course etiquette

We are expecting a large number of participants in this session. We suggest everyone follows [these few simple rules](#) for the course to run as smoothly as possible.

Questions

If you have any questions/problems that you would like to share and are applicable to the whole class please write them below. A tutor will answer your question.

Write your question after the last one you can see in this document and write your name.

Day 1

1. **<Max Beesley><Per sequence GC content Bad Data - How come the C's and G's are not equal? I understand its bacterial contamination. The two lines, do they represent G, and the other C?**

Ash: The "GC content" is the percentage of a sequence that is made up of G or C, i.e. the ratio of [GC]/[AT]. The blue line on plot is the inferred ideal distribution, the red line is the actual from the data.

2. **<Max Beesley><Where does FASTQC have to be installed for this to work?>**

Ash: There are installation instructions on the FASTQC website for Windows, MAC, and Linux. It's actually a java app, so you also need to have java installed.

<Max Beesley> I have it installed as a stand alone program but the command fastqc does not work on command line, I am unsure how to integrate it?

Ash: What sort of computer are you using (Windows/Mac/Linux)?

<Max> Mac

Ash: To get the command line version download the zip file marked Win/Linux and unzip it. In the unzipped directory there is the `fastqc` file. You'll need to add execute permission by running `chmod +x fastqc`.

3. **<Melinda>****<I removed both html and zip files and tried running fastqc -o fastq/MCL1.DL.fastq.gz but it doesn't run FastQ again?>**

Ash: You need to specify a directory after the '-o' for the report to be written to: `fastqc -o fastq fastq/MCL1.DL.fastq.gz` will write the report into the fastq directory

4. **<Ben>****<I was behind earlier so missed how to open the HTML report in the browser>**

Ash: You can either use the file manager on the desktop - at the bottom next to the terminal icon - or type `xdg-open` followed by the report file name in the terminal.

Thank you!!

Ash: Please use the raise hand and chat for technical issues and use this document for more theoretical questions.

5. **<Raina>****<The "Sequence Duplication Levels" in the FastQC report does not look good. It says only 48% of the sequences will remain after deduplication - do we need to address this? If so, how? >**

Ash: With Single End sequencing we expect high duplication rates as the transcriptome is relatively small and we have many millions of reads and so there is a good chance that we will get two identical fragments that are from different original mRNA molecules. This level of duplication is not an issue for mRNAseq,

6. **<Kate>****<Why is the 'Per base sequence content' marked with a red X, does this denote it is bad, as it looks fairly similar to the good example in the powerpoint examples?>**

<Jon>**<FastQC was originally designed for DNA Sequencing, therefore when things are marked red sometimes this is just because it is not right for DNA Sequencing. It is better to look at the actual plot to determine if it looks O.K.>**

7. **<Lorea>****<Are we going to learn how to do read trimming?>**

<Jon> We will not be doing trimming of the reads during this course, however if you need to do read trimming with your data there are two main programs for this: Cutadapt and Trimmomatic. Both are easy to download, install and use. If you have any problems with this you can speak to somebody as this is a very common step in the analysis.

Ash: In the "Extended Materials" there is an exercise that shows how to use Trimmomatic

<Lorea><I assume then that if the quality of the fastq file is good, then read trimming is not compulsory. I'm asking this because some people says that reads come with adapter sequences and then you should remove this before alignment. But others say that in most sequencing facilities they remove this already and you should not have this sequence in the fastq file you receive

Ash: If the data is of good quality there should be little or no adapter contamination in the reads. The problem only occurs when the fragment that is being sequenced is shorter than the read length - we will look at this later. Small amounts of adapter contamination are not really a problem for alignment as the aligners can cope with this. Many people do carry out trimming as a standard step of the workflow, but with the generally high quality of sequencing these days and the way aligners can cope with small amounts of contamination, it is not usually necessary.

8. **<Carlos Bueno><What is the best alignment tool for to be used when working with paraffin-embedded formalin-fixed samples?>**

Ash: The choice of aligner is more to do with the type of sequencing you are dealing with - RNAseq requires gapped alignment but doesn't need precise mapping, DNA alignment for variant calling etc. requires more precise mapping. I haven't found that we needed a specific aligner when working with data from paraffin-embedded formalin-fixed samples.

9. **<Kate><Do you need to convert your fastq file to a SAM file?>**

<Tutor>SAM is the output, FASTQ the input

<Kate>So the SAM output is the fastqc.zip file that was generated alongside the html?

<Tutor>The FASTQ file is the input of the HISAT2, which outputs SAM. The first step we did was just quality control and does not modify the file at all

10. **<Natalie van Dis><How can you check beforehand if you have enough space for your analysis (if SAM files take up a lot of space)?>**

<Jon>< This is a hard question, the size of SAM files depend on the number of reads you have, typically 30 Million reads will be around 15GB each depending on the size of the reads and a few other factors. On a UNIX machine you can use the command "df -kh" to check available space.>

<Natalie van Dis><And the 15GB is for PE with 30M reads? Or SE? And do you mean 15GB per read or per sample?>

<Jon> Per sample, this is only approximate, I will have a look through some of my SAM files and get a more accurate number for you. Bear in mind that after converting SAM to BAM the size is reduced by ~10 fold.

11. **<Lorea><I used STAR in the past to align my Fastq files, where I also need to generate the index first. In STAR you need to specify the length of the index, i.e. 50, 100nt, etc, which should be the closest to the length of your reads. This means that I need to create indexes every time I analyse fastq files of different read lengths. Is this right? Alternatively, can I use an index of 50nt to align reads of 100nt or the other way around? >**

Ash: You should try to use an index that has been generated for the length of the read (or close to) that you are aligning. This is to do with how much overlap across exon junctions is allowed. If you have 50 and 100 nt reads in different files you should use the appropriate index. Of course, once you have the index for that read length you won't need to generate it again if you get new data of the same length.

<Lorea> **<Thanks for the explanation. When running Hisat2, how do we choose the length of our indexes in the exercise? I don't see anything in the command which determines the index length>**

Ash: The hisat2 index is different to the way the STAR index works and doesn't need the read length.;!/
</

12. <Noe><I cannot access the directories as you are indicating on the exercise. I need to put the whole directory address. Is that normal?>

<Tutor> You can either use the full directory address (Full path). Or you use a relative address (relative path) this is the location of the desired place compared to where you are. So use commands like "pwd" and "ls" to find out where the directory is that you are trying to reach.. As for the exercise, the commands should work as they are displayed, just try and make sure you are in the correct place when you start.

13. <Eve><I'm getting an error that says hisat2-align exited with value 1. This is my code: hisat2 -p/-- threads 7 -t/--time -x references/hisat2_index/mmu.GRCm38 {-U fastq/MCL1.DL.fastq.gz} [-S bam/MCL1.DL.sam] >

<JoN>

A couple of things here: When copying from the usage statements you do not need to include the brackets <> {} []. Of note here is that the brackets do have different meanings when reading the usage statements - <> Compulsary inputs, [] optional inputs., { option 1 | option 2 } - This says that you must choose either option 1 or option 2.

Further to this when you have a parameter like this in the help: -p/-- threads. The slash denotes an option. You can either use -p 7 OR --threads 7.

hisat2 -p 7 -x references/hisat2_index/mmu.GRCm38 -U fastq/MCL1.DL.fastq.gz -S bam/MCL1.DL.sam

14. <Marie><What is the meaning of FLAG 272? I can't find it in the table, but I have it in one of my lines>

<Zeynep> I will explain it when we come back, but you can use this tool to decipher SAM flags: <https://broadinstitute.github.io/picard/explain-flags.html>

272 means the read is mapping to reverse strand and it's not a primary alignment (meaning that it aligns equally well to several locations in the genome).

15. <Jagon><If the interface shows "Time loading references: 00:00:00", does it mean its working or not? This is after i typed in my code for Exercise 2; But then the exercise said it will take 5-10 min? Ah alright then , thanks!>

Ash: It just means that it took less than a second to load the reference index into memory. It's worked fine. I think these machines are a bit more powerful than the ones we used when I wrote the materials :)

16. **<Natalie van Dis><Field 9 in the SAM/BAM files gives Template length, is this only for paired end? >**

<Zeynep> Yes, fields 7 to 9 are reserved for paired end alignments.

<Natalie van Dis><And are the only values it can take 0 1 or 2 ?>

Ash: No, the Template Length is the length of the mRNA fragment that was bound between the adapters. Although we are getting 50 nt reads, the actual fragment can be much longer - we only sequence the first 50 bases. With paired end reads we can infer this length based on the distance between the places the two reads align, with single end we can't know this.

17. **<Kate><Whilst the data we are using here did not require QC, do you have any guides/references on QC-ing raw reads, i.e. QC before the hisat2/sam/bam stage?>**

Ash: The QC of the raw reads was carried out in the first section using FASTQC, this gave us the necessary information about sequencing quality, GC content and adapter contamination to assure us the sequenced data is of adequate quality.

<Kate>Yes, but what if you decide at this stage, you need to remove some samples due to poor quality/etc? Or is this not done with RNA-seq data?

Ash: Each sample would have a separate FASTQ file, so if some of your samples are of poor quality you would just not use that FASTQ.

<Kate>Ah, ok - I'm used to working with genotype data in which you QC everything together and apply filters to the overall dataset, this makes sense here, thank you.

Ash: We are just working with one file here to save time, normally you would look at QC for all the samples

<Kate> So back to my original question, when you run fastQC for all your samples, how would you then filter out 'bad data'? Or is it just the case of removing certain fastq files from your overall data, not necessarily applying filters (i.e. call rate/MAF/etc) as you would do for genotype data?

18. **<Yifan><How do you decide how many cores to use with -@ ?? Or do we always use 7?>**

Ash: In this case our machines have 8 cores, we need to leave one for the operating system display etc. and so we use 7. If you are on a cluster then you can use as many cores as are available, however, there is a diminishing return in terms of speed as you increase the number of threads. For alignment 8 threads is about optimum.

19. **<Jagon><How do you view the sorted bam and since its a binary file, what does it represent? Thanks!>**

<Tutor><You need to use "samtools view <x.bam>" to view the binary file. The purpose of a bam file is just to make sam files smaller and aids the creation of indexes. It contains exactly the same information as the SAM file.>

20. <Zeinab><can you explain again how we can access the sam data details as you've shown in the practice session>

Zeynep: You can use *less* command to view a text file in linux, SAM format is in plain text format, so you can use *less* to see it.

21. <Natalie van Dis><So is bias because of PCR a problem for RNAseq and if so, how can you distinguish between PCR bias and the duplication you expect with RNAseq analysis?>

Ash: If you are referring to the bias in the coverage plot, this is due to RNA degradation. The RNA is degraded by RNAases from 5' to 3'. So, if we have degraded RNA then there is more tendency to extract fragments from the 3' end of the transcripts. This will be bias worse for shorter transcripts.

With regard to distinguishing PCR duplicates from true duplicates, the only way to do this currently is to include Unique Molecular Identifiers (UMIs) in your library prep. These tag each molecule with a random nt barcode, which we can sequence. The chance of getting 2 true duplicates with the same UMI is astronomically small, so if we have a duplicate with the same UMI we can assume it is a PCR duplicate and discard. In practice, duplicates tend to make little difference to the downstream differential expression analysis and the current cost of including UMIs means that this is not regularly done for RNAseq at the moment.

Natalie van Dis<Ok, so the high duplication rates seen both in the raw QC and alignment QC are expected and can just be ignored for RNAseq data?>

Ash: Yes, duplication rates up to 60 or even 70% are not unusual and can be ignored - for single end. With paired end the duplication rates tend to be lower as the chance of getting two identical pairs is much smaller.

22. <Phoebe><My alignment rate with samtools flagstat has come out as only 11.8%, do you know where I might have gone wrong with this?>

Ash : answering in chat...

23. <Wei><sorry if I've missed this but what does all the different .bam files mean in the metrics folder? Do they represent the rna-seq data being broken up into different files? Or different experiments?>

<Jon><Yes different samples, so we just worked with one sample today (for speed and ease) but we will be working with more samples for the differential expression>

24. <Name><Can we have the code for 2.1 and 2.2 pls

Ash: All the solutions will be posted on the website shortly.

25. <Carlos><What is the explanation for the presence of so high percentages of duplicates?>

<Karsten>In RNA-Seq a high duplicate rate is expected as each RNA molecule will be present in multiple (100s-1000s) of copies

26. <Natalie van Dis><A colleague told me that for nonmodal organisms you might have to adjust for how much nucleotide diversity you will allow during the mapping process. Do you have any guidelines for this? And how can we specify that for hisat2?>

<Jon>< I have not heard of this for a mapping parameter but I will look into it.>

<Natalie van Dis><Thank you! My colleague said that you expect much more nucleotide diversity for wild animals compared to humans, so that you might have to try out allowing for different levels of nucleotide diversity and see what matches the best. But would be great if you have some more guidelines.>

<Jon> Ahh yes sorry this is my fault, I didn't read the question carefully enough. So yes, when mapping reads to a reference genome it is important to look at the "mismatch rate" and then the total number of reads mapping to the reference genome. If you have a diverse species and you have a high mismatch rate and/or a low mapping rate you can adjust the --score-min parameter to allow reads to map to the genome with more mismatches than are allowed by default. However you should be careful with this parameter as it may cause erroneous mapping. If your sample is very different to the reference you may want to speak to your PI or other person to have a look into creating a reference genome for your sub-species or cultivar.

27. <Yifan><I'm still a bit confused how to check the quality of the RNAseqmetrics read alignment? Is there a threshold or the percentage of at least how many reads have to be mapped to the coding region to be accepted? >

Zeynep: This will depend on the organism you're working on and the quality of the sample (library prep, RNA quality). With well annotated genomes (human & mouse) and *good* samples, we observe this ratio be around at least 80-85%. If you have lower ratios than this, you might need to investigate further. These are diagnostic tools and they will only indicate that there might be issues with your data, but it will be up to you to investigate further and decide if you need to (or can afford to) exclude this sample or not.

28. <Noe><How can you tell apart technical duplicates from actual increase in RNA/gene expression?>

<Karsten>The only way to accurately identify PCR duplicates in RNA-Seq is to label the mRNA/cDNA molecules prior to PCR amplification with unique identifiers called UMIs (unique molecular identifier). You would then be able to identify PCR duplicates as reads with the same UMI. Some protocols include this in the library preparation. In general this decreases the "technical noise" in your data, and is a popular approach in scRNAseq.

29. <Natalie van Dis><How can we visually check the mapping using the .bai files?>

<Jon><You can visually view the alignments with IGV - This program takes bam files and a reference genome as input <http://software.broadinstitute.org/software/igv/>>

30. **<Raina><Please could you explain again what does the indexing do? Why do we need to index the reference file and the bam file? Also why do we have the “.bam.bai” file format?>**
<Zeynep> Indexing enables rapid and accurate alignment of short reads (~100bp and less) to large genomes (~3 Gbp). Without indexing, the alignment step becomes computationally very expensive (takes too long, requires large memory). Indexing the bam file also has the same benefit: the tools that use these bam files will work much faster with index files.
31. **<Bingnan><In the GTF files, what is the meaning of the 6th column? Why for all records there is a dot?>**
<Jon><The 6th column is a score, it is only applicable to GFF files, GTF files still have this column because they are a type of Gene Feature Format file >
32. **<Lorea><In exercise 1, why are there ~3-4x CDS than transcripts? I would assume this number should be similar>**
<Jon><Yes so in the annotation there is one CDS annotation per exon, not per transcript. So a coding transcript with 4 coding exons will have 4 CDS lines associated.>
33. **<Kate><Going back to the first session this morning, how would you run fastqc on multiple files?>**
<Jon>< Fastqc accepts “wildcards” so you can run:
fastqc fastq/*.fastq.gz (there is also a -T option for threads) >
<Kate> Thank you, I was just reading the wildcard section of ‘4. Running featureCounts on multiple samples’ and was wondering if the same could be applied to fastqc.
34. **<Natalie van Dis><I’ve read that you can either do counts at the gene level or at the transcript level. What’s the difference and when would choose to look at transcript level rather than gene level?>**
<Karsten>This depends on the question you’re interested in answering. In most cases you’d be interested in identifying differentially expressed genes because you’re most likely trying to identify pathways/genes that are more/less expressed in one condition versus another. However, in some cases you might be more interested in differential exon usage. In the latter case you would want to summarize the counts by transcript/exon level rather than gene level.
<Natalie van Dis><And what if your reference genome quality might be not so good, would that also be a reason to look at transcription level?>
<Karsten>Hmm I don’t have much experience with non-model organisms so maybe someone might correct me but I’d expect that if the genome annotation is poor that the transcript level annotation would be if anything worse than the gene level annotation. So I don’t think this would be a reason to count per exon/transcript rather than per gene.

35. **<Lorea><When running `head counts/MCL1.DL.featureCounts`, is it ok to interpret that gene with id `ENSMUSG00000102693` (first line) has only one exon, whereas gene with id `ENSMUSG00000051951` (third line) has 7 exons?>**
<Tutor><The transcript `ENSMUST00000193812` has one exon and the transcript `ENSMUST00000162879` has 2 exons >
36. **<Name><i have to leave the course early today. Are tomorrow's sessions dependent on today's exercises?>**
Zeynep: Tomorrow we will continue with differential expression analysis - but we will provide the necessary files (so it doesn't depend on the output of today). Recording of today will be available later in the evening, so you can catch up on the theory if you want.
37. **<Lorea><When running `head counts/MCL1.DL.featureCounts`, where will it show the reads which map to two different exons because they are split reads due to splicing?>**
<Tutor><This is not shown with feature counts, you will need to interrogate the BAM / SAM file to pull out reads that have a gap in their alignment.
<Lorea><will they be part of the unassigned reads then?>
<Jon> split reads that map across exon junctions will be counted my feature counts. Depending on the parameters these reads will be counted as a read for that gene / transcript.
38. **<Rob Horne><Would it be possible to get brief instructions on how to set up this environment with all the associated packages? Will this only work on ubuntu?>**
Zeynep: Most (if not all) of these tools work only in linux. For creating your computing environment in a controlled way, I can recommend miniconda: <https://docs.conda.io/en/latest/miniconda.html> You can install the tools we have used from the Bioconda channel: <https://bioconda.github.io/user/install.html>
<Rob Horne><Ah OK would it be possible to run this on mac with Linux running in VirtualBox?>
Zeynep: You don't need a virtualBox, these tools will work in Mac (as macOS is a unix system).
39. **<Lorea><If there is a read that maps to a gene which is within another longer gene in the same strand, for example a miRNA gene which is within another longer gene, is there anyway to know if a given read belongs to the short or the long gene? Not sure if I explain myself clear>**
Ash: Perfectly clear. This is not really possible if both the genes coding regions overlap as there is no information in the sequence that tells us which meta molecule the read comes from. If, for example a miRNA originates in the intron of another then yes, you can figure this out. If this is the sort of thing you are interested in then you may need to look at alternative sequencing strategies that will allow you to sequence full length transcripts e.g. nanopore.

40. <Bingnan><What is the difference between ‘feature’ and ‘attribute’?>
Ash: Have a look back at the slides. Each row in the gtf is a feature, this may be an exon, a gene, a transcript etc. The attributes are the annotations of the features, e.g. each exon belongs to a particular gene and so has a gene_id attribute.
41. <Jagon><I encountered this msg “ ERROR: no features were loaded in format GTF. The annotation format can be specified by the '-F' option, and the required feature type can be specified by the '-t' option. The program has to terminate. What does this mean?
\$ featureCounts -t gene_biotype -g gene_id --primary -a references/Mus_musculus.GRCm38.97.gtf -o counts/MCL1.DL.gene_biotype.featureCount bam/MCL1.DL.sorted.bam
Is there a problem with my code? I run the code w/o and it still gave me that error
Yeap thanks!
<Jon> I am guessing you want to count all the reads mapping to exons for each gene_biotype?
featureCounts -t exon -g gene_biotype --primary -a references/Mus_musculus.GRCm38.97.gtf -o counts/MCL1.DL.gene_biotype.featureCount bam/MCL1.DL.sorted.bam
42. <Noe><Why did we use the .bam files instead of the bam.bai files?>
<Jon><BAM files contain the data (reads mapped to the genome). The bam.bai files do not contain data, only the indexes of the bam files, these bam.bai files allow other programs to efficiently search the BAM files. >
43. <Johan><What tools are there available to count towards splice variants instead of against genes?
<Karsten>You can do that with featureCounts. You can specify the level at which you want to summarize your counts when calling featureCounts. If you are interested in the number of reads mapping to exons rather than genes (that would be the case if you’re comparing splice variants for example) you need to set -t exon -g exon_id. The -g flag essentially specifies at which level the counts are summarized.
44. <Natalie van Dis><I read the Tuxedo protocol where they say hisat2 output is streamlined to be used by Stringtie to do counting. Any reason to prefer Subread over Stringtie? Or is either tool ok to use? >
Ash: The Tuxedo workflow is well established and there is a python tool that they provide for generating the necessary gene expression matrix from the Stringtie output for input into DESeq2 or edgeR for differential expression analysis. The main difference between simpler methods such as featureCounts or HTseq and StringTie is that StringTie also does de novo assembly from your reads to discover novel transcripts. This is very useful if the organism that you are working with does not have a well established transcriptome reference, or if you are specifically looking for novel isoforms or maybe fusion gene products, however, it may just be introducing undesirable noise to your analysis.

45. **<leila ><form a bam file or the fastq file how we can know if the sequencing was paired end or not?>**

Ash: For the raw reads, the fastq files for paired end come in to forms. Most commonly you will get a pair of fastq files for each sample, one for read 1 and one for read2. The order of the reads in the two file will be the same. Alternatively, you may get an “interleaved” fastq (this is not so common any more) where the read 1 and read 2 for each pair alternate through the file.

For the bam file, you can just run ``samtools flagstat <bamFile>`` and it will tell you how many paired reads there - the aligner adds tags to the alignment to identify the pairs. This will be 0 if the data are single end.

Day 2

46. **<Rob Horne><Is there any way to save the work we’re doing on RStudio to our own computers for later reference?>**

Ash: If you save your script on the machine, you can then download it using the Files panel. Select the file and then click the “More” drop down, it has an option to “export” to your local machine. Stephane will demonstrate this at the end of the session.

Many thanks!

47. **<Max Beesley><Can we download the original fastq file onto our own computer? I would like to run through all this again at the weekedn>**

Ash: In the extended materials there are instruction on how to retrieve all the original raw fastq from the GEO repository. However, you will not be able to do the alignment on your laptop/desktop as it is too computationally intensive. You are welcome to download the featureCounts output if you just want to run through the R parts of the course. Also, the virtual machines will be available to you over the weekend.

48. **<Angie><What’s the difference between using "Geneid" and ‘Geneid’?>**

Ash: None at all, you can use `""` or `””` effectively interchangeably in R. Personal preference really.

49. **<Munise><Why did we use 5 to filter the count matrix row sums in R? Is that just an example?>**

Ash: Yes, using 5 reads is a little arbitrary. For the purposes of the differential expression analysis we don’t actually need to filter out genes - we will see tomorrow that DESeq2 does this for us. However, it is easier to remove genes that are definitely not informative for the purposes of plotting and clustering. As you saw there were 30K genes with no reads or so few reads that this was probably just noise. They would squew all of our plotting and make it hard to visualise patterns in the data. With regard to the choice of 5 reads, if, for example, you had more samples, you might want to increase this number. Generally, I use a rule of thumb that I am aiming, in mouse or human, to filter down to about 20-25 K genes.

50. **<Ben><Is there a reason we transform the data after taking out the low counts? Seems like doing it this way round skews the final distribution as we've removed lowly expressed genes before transforming>**

<Jon><Low counts have a vary high variability, just because of their nature, e.g for gene X in sample 1 you have 1 read and sample 2, 2 reads. This is a 2 fold difference and clearly would be not very informative. Then the transformation of the rest of the counts is to address the issue of heteroskedasticity >

Ash: Also, with genes that have just 1 or 2 reads in a couple of samples it is much more likely that these reads are just technical noise rather than actual expression of the gene. Note that we still have >20K genes left.

51. **<Tsveta><How do we choose the best way to transform our data?>**

<Jon><Maybe there is a more encompassing answer out there, but these 3 types are being shown just to highlight the differences. Generally rlog is considered the best. However it maybe slow with A LOT of samples, in this case VST is faster.

52. **<Carolin> How do you decide which type of normalisation is most appropriate for your dataset? Which normalisation do you most commonly use?**

<Jon><Normalisation allows you to compare across different samples and between genes. The two most commonly used are RPKM and TPM normalisation. TPM (Transcripts per million) is the most robust as it accounts for difference in transcript length and not just gene length. So If you are looking to plot say the expression of Gene X across your samples (in a barplot of similar, best to plot TPM.) >

Ash: Also we only use the normalised data for things like visualisation and clustering, the differential expression analysis works with the raw counts. Never feed edgeR or DESeq normalised data such as TPM.

53. **<Carolin><Thank you! That makes sense! I am not sure this is within the scope of this course - but I have used voom transformation/normalisation in the past, and was wondering what the main differences/benefits of one over the other were? Do you have any thoughts/opinions on voom?>**

54. **<Zhaotao>When we call the log2 command, I understand we added +1 to the 'countdata' to remove the 0s, do we not need to do this with rlog as well?**

Ash: No, rlog handles this internally.

55. **<Maria><How can we get a copy of the R-script for ourselves and also download the course materials?>**

Zeynep: The code will be available in the website and the source materials are available via github:

https://github.com/bioinformatics-core-shared-training/RNAseq_November_2020_renote

56. **<Jagon><How do i save my own R script?>**

Ash: Save via File >> Save. To download see Q46.

57. <Maria><Zeynep, but what about our R-script with all the comments?>

Ash: See above and Q46.

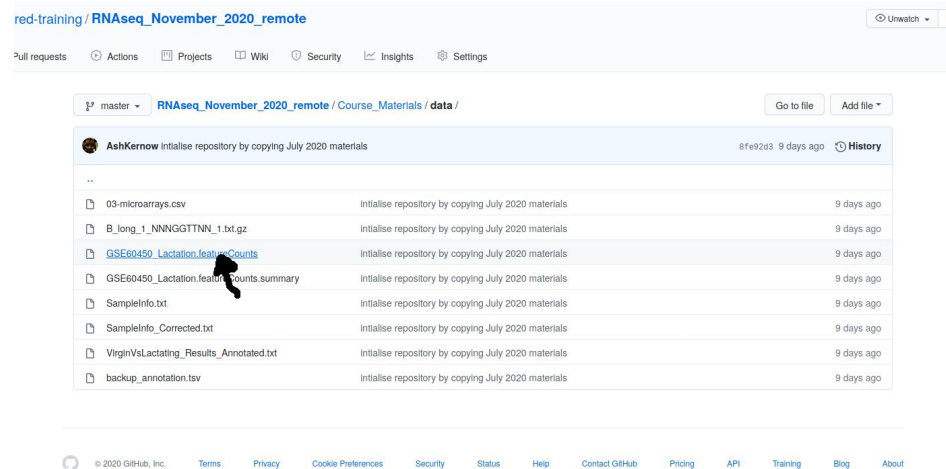
58. <Lorea><how can i save the GSE60450_Lactation.featureCounts and SampleInfo.txt files in my computer? I tried to go to GEO here

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60450> but the table which I downloaded as GSE60450_Lactation-GenewiseCounts.txt.gz does not look the same>

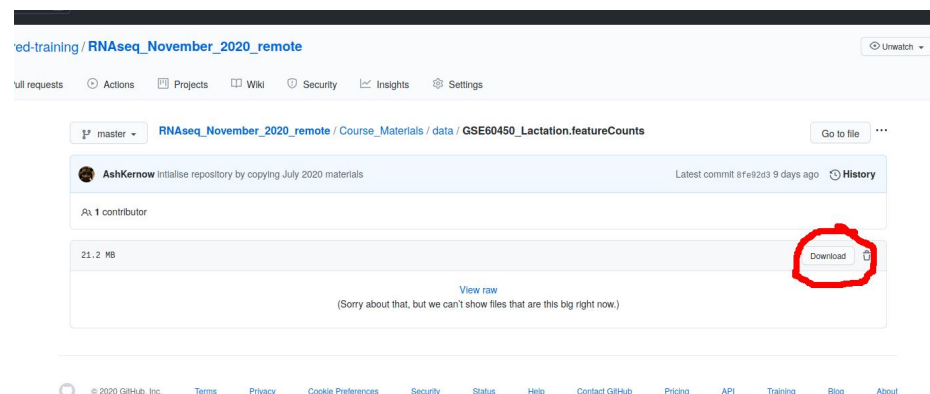
Ash: No, the table we are using was generated using featureCounts from the aligned bams as we did yesterday in the last session (although yesterday we only did chr15 to save time). You can get all the course materials for the R parts of the course from our GitHub, the link is on the webpage or Q55 above.

<Lorea> ok, but I go to the GitHub link above Course_Materials/data and I right-click on the file [SampleInfo.txt](#) or [GSE60450_Lactation.featureCounts](#) and download linked file. But it downloads an .html file. How to download the actual file?

Ash: Don't right click the file name, click it and it will take you to another page, where you should see a download button:



The screenshot shows a GitHub repository page for 'red-training / RNAseq_November_2020_remote'. The file list is displayed under the path 'Course_Materials / data /'. The file 'GSE60450_Lactation.featureCounts' is highlighted with a black mouse cursor. Other files in the list include '03-microarrays.csv', 'B_long_1_NNNGGTTNN_1.txt.gz', 'GSE60450_Lactation.featureCounts.summary', 'SampleInfo.txt', 'SampleInfo_Corrected.txt', 'VirginVsLactating_Results_Annotated.txt', and 'backup_annotation.tsv'. The repository was initialized by 'AshKernow' 9 days ago.



The screenshot shows the view of the file 'GSE60450_Lactation.featureCounts' in the GitHub repository. The file size is 21.2 MB. A red circle highlights the 'Download' button in the top right corner of the file view area. The repository was initialized by 'AshKernow' 9 days ago, and the latest commit was 8fe92d3, also 9 days ago. The page includes a 'View raw' link and a message: '(Sorry about that, but we can't show files that are this big right now.)'.

<Lorea> Yes, I've tried that, but it opens the file in a new tab and again I don't know how to save that into my computer.

https://raw.githubusercontent.com/bioinformatics-core-shared-training/RNAseq_November_2020_remote/master/Course_Materials/data/GSE60450_Lactation.featureCounts

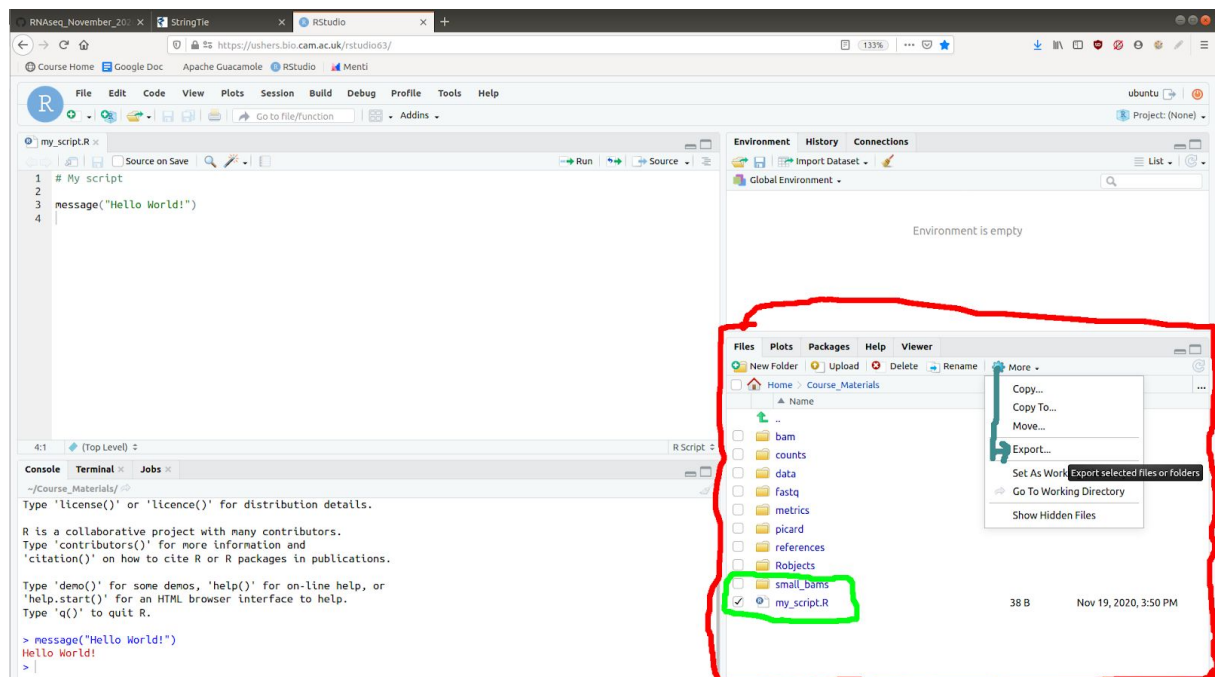
<Karsten> You can then select "Save As" under "File" in your browser menu.

<Lorea> this option "save as" is not available. Don't worry, I'll try later and if not I'll ask again.

<Lorea> <I managed now following the instructions given in Q46, rather than GitHub. Not sure why I couldn't from GitHub, but anyway I managed now. Thank you!>

59. <Maria><in Q46, when you say Files, is this in R?, Sorry, I can't find it> FOUND IT NOW

Ash: In RStudio there are 4 panes, the bottom left (usually) has the "Files" panel, there you can select a file and then click "More" in the menu bar to find the option to export it.



60. <Noe><Maybe we will go back to that tomorrow, but I was wondering what happens with duplicates? Is DESeq2 removing them? Should we remove them or should we take them into account? How these affects the analysis?

Thanks!>

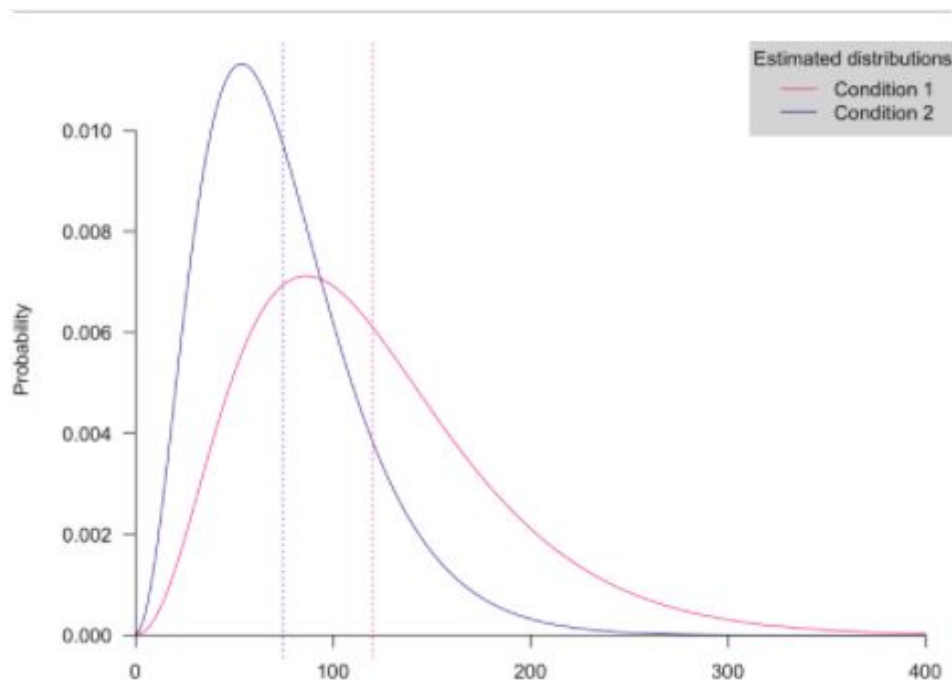
Ash: You have to handle duplicates during the counting step that we did last thing yesterday. You can choose to exclude them, however, given that we are generating maybe 20 million reads from a transcriptome of maybe 30-50 MB of which only a subset of genes are expressed, we actually expect to see a large number of genuine duplicate reads that originated from different transcript molecules. We would not want to discard these as they are a function of gene expression. In general we have found that including duplicates has little influence on the differential expression analysis.

With paired end data, the likelihood of getting an identical pair of reads from different molecules is much smaller, so you can more reliably discard duplicates (there is an option in featureCounts).

61. <Natalie van Dis><Maybe I missed it, but do we need to check the DESeq2 estimated distributions for the different conditions and what does it tell us exactly? What should we pay attention to?>

Ash: Sorry Natalie, I am not entirely clear what it is you are asking, but tomorrow, we will actually go through the practical application of all of this using DESeq2, so hopefully this will all become clearer.

<Natalie van Dis> So Dominique showed how to plot the estimated distributions for the different conditions you are modelling, and I was just wondering what to do with those plots once you made them. So should you check them to see if your model fit is ok, or does it say something about the differences between your conditions, or what should we look for? I mean this plot:



Dom: It was just an illustration of the assumed distribution of the counts for a given gene (#1) per condition. There is nothing much to do... it is just for you to see the estimated/fitted distribution: in condition 1 (reference), the counts have an asymmetrical distribution with most values on the interval [10; 300] while, for condition 2, the counts have an asymmetrical distribution with most values [5; 200].

62. <Noe><Regarding FDR, then we need to pay attention to the adjusted p value rather than the p value, don't we? In Dominique's examples some genes had significant p values (e.g. <0.05) that were not significant anymore after FDR.>

Ash: Yes, this is correct, you have to use the adjusted p-value to determine the set of genes you consider to be statistically significant.

63. <Noe><Following question 60, does this mean that sometimes single end is better than paired end when we do not need to account for transcript variants? Or that it has higher depth? Thinking about what to chose for experimental design when deciding how to better analyse samples. Sorry if the answer should be obvious from yesterday's lecture>

Ash: Sorry, no, for sure paired end is better, in this case we can discard duplicates and get more accurate gene expression measures. However, in practical terms, there is little impact in downstream gene expression analysis. With regard to the SE/PE decision, in practice you may find that it is a moot point as Illumina have long been pushing us towards doing PE for everything and the newer library preps a nearly always PE.

64. <Yifan Wang><I saw the term "FPKM (short for the expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced)" in the analysis report from the sequencing company. They used this FPKM to do the quality check for gene expression level analysis and did log₂ transformation distribution with FPKM values. I'm a bit confused. What's the difference between FPKM and the countdata we talked about in the morning lecture? Are they the same?>

Ash: No they are not the same. FPKM stands for "Fragments Per 1000 base pairs Per Million reads". These are normalised counts. These are normalised to gene length (Per 1000 bases in gene length) and library size (Per Million reads in the library).. There are three methods that you might see for this - RPKM (not recommended any more), FPKM and TPM (recommended). The methods are straightforward, but the differences between them are subtle but important. It's a bit too long to get into here so I will direct you to this blog post for starters: <https://rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/>

65. <Yifan Wang><I also saw people do correlation coefficient of samples between groups analysis as part of the quality check, and we >

Ash: Yes, this another good way to check that your replicates look more similar to one another than to samples in other sample groups. We haven't included that in the course, but it is worth doing.

<Yifan Wang> Thank you!! I was wondering what's the difference between correlation analysis and the PCA analysis, the final result feels quite similar to me.

Ash: They are very different. Correlation - pearson correlation usually - is looking at the distance between different samples in n dimensional space, where n is the number of genes you use to calculate the correlation coefficient. PCA is looking at that same n-dimensional space, but trying to rotate the axes so that we get the most variance explained in the least number of dimensions. This is a very short explanation, it is definitely worth doing some extra reading about PCA, it is useful in many ways other than just looking at the relationships between samples and sample groups.

66. <Noe><In order to try to account for variance in samples, would it be a good option to pool random equivalent samples experimentally, even if at the end there are only 3-4 biological replicates? This is, if you can only afford 3-4 samples per treatment (WT vs Mutant in the exercise) making every sample a pool of three fetuses from different mothers may help to get rid of technical noise (collecting the fetuses at day 15 +/- 12 hours)>

Ash: This would really have to be decided on a case by case basis depending on the experiment, but generally, especially with mice, we would say no. We want to measure the variance due to individual variations and pooling samples would mask this. The noise between fetuses from different mothers is biological rather than technical (there might also be technical variation due to e.g. the different mothers were fed differently).

67. <Noe><Does it make sense having different sample size for different groups? E.g. 3 WT & 5 Mutant vs 4 WT & 4 Mutants?>

Ash: This is not a problem. If for some reason you were to lose two of your WT samples, it is still better to keep all 5 of your experimental samples. If you expect one group to have higher biological variance, e.g. you are giving a treatment to which different individuals have different responses, you may want to increase the sample size in that group. Again, *ideally* all of our sample groups would be the same size, but practically this is not always possible.

68. <Max><Where can we access the annotated notes/slides? Thanks>

Ash: All of the materials are available on the course website linked at the top of this page.

Day 3

69. <Name><In each folder are we? There is nothing in my R objects. It seems all my folders are empty in my terminal > Okay. I am doing it. Now it is working. Thanks :D

Tutor<Try file --> quit RStudio and reload it>

70. <JAgon><Can someone explain the ~ function again? Thanks!>

Ash: When used in a formula the tilde basically means that the thing on the left is a function of the thing on the right. You can think of it as equivalent to “=” in a formula such as $y = mx + c$. In our case we are saying “Gene expression ~ Status”, meaning in our model, gene expression is a function of, i.e. is affected by, the status of the mouse. We don’t need to actually specify “Gene expression” in this case. Thanks ;)

71. <Max><How do we know which columns we want when altering the modelmatrix? Is the Virgin status now the intercept column?>

Ash: Yes, virgin is now the intercept, you can tell this because it is the one that is missing in the column headers of the model matrix.

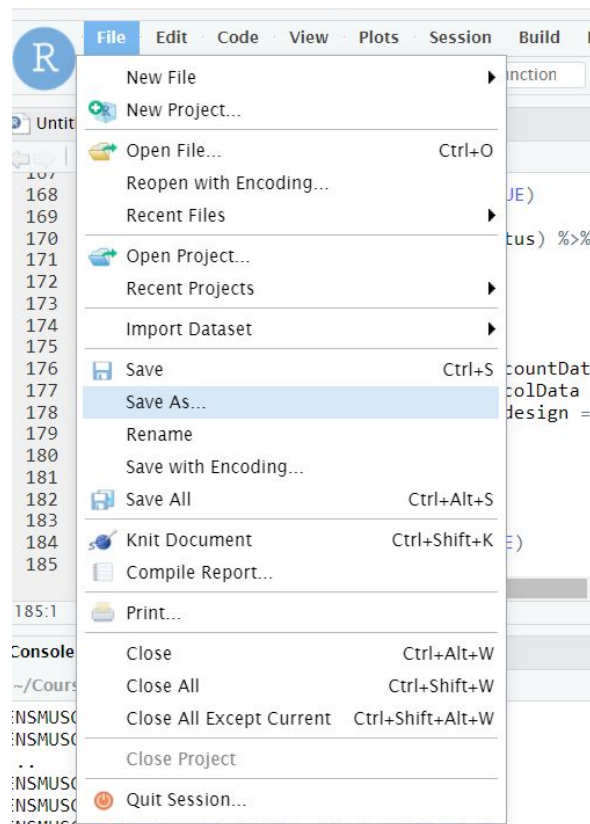
Max And that is because virgin is the column of interest?

Ash: You want to set the intercept to be the “control”. When we get the differential gene expression, the fold change will be relative to the intercept. In real terms it doesn’t actually matter what the intercept is, you can always extract the contrast that you want, no matter which factor is the intercept. It just makes the interpretations simpler if you set the “control” to be the intercept.

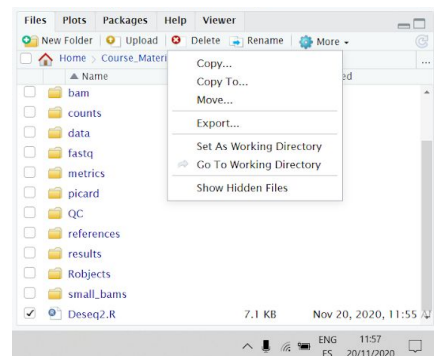
72. **<Jagon>****<Sorry, but i've asked this question before but i have not figured out how to download my code into my own laptop, thanks (can't seem to find that “More” option)>**

<Abbi>**<it only appears if you are in the files tab, that confused me the first time too>**

Jargon: You mean this?



<Melinda>**<I did it like this: clicked on the script and then More>Export! Thanks :)**



73. <Natalie van Dis><For a time series experiment, I guess the model would be something like [~ treatment * time point]. How does time point need to be coded? Also as a factor with levels in the correct order or as something else?>

Ash: Hi Natalie, yes that is the correct model. The consideration here is whether you want to treat time as a continuous vector or discrete intervals. If you are expecting a time ~ response curve in gene expression then you can consider it to be continuous and actually have it as a numeric vector. The model will then try to fit a linear change in gene expression against time. This doesn't seem to be appropriate in most cases and so most of the time series that I have worked on, we have treated time as discrete. There are then various possible avenues for analysing the modelled results. Rather than go into them here I would direct you at the DESeq2 workflow document, which has a small section on time series analysis. After that you can look deeper at various posts on the Bioconductor forum.

<Natalie van Dis><Thank you! How do you decide whether continuous would be appropriate or not? Do you fit the model and then look at model fit?>

Ash: No, in this case you should make your decision based on the biology you expect, do you expect your genes to be responding with an increase/decrease in expression that is linearly related to time, or do you expect them to go up and down at different time points. Remember (Abbi will talk about this later) you can only apply one model to ALL of your genes. This applies if you are going to use the current workflow for your analysis. There are other tools for time series analysis that may be appropriate, but I am afraid I have no experience with using them. The Bioconductor community is your friend in this case.

74. <Munise><Would you remove rows with padj == NA at the stage of getting the results out from DESeq?>

Ash: Yes, for the downstream considerations (this afternoon) we just remove genes with NA in the padj column.

75. <Erika><The results that are positive in the log2foldchange column of your analysis are negative in mine, and your negatives are positives in mine. The numbers are the same, just +/- differences. Do you know what I did wrong? > You were right. I forgot to factor the dataset using virgin as the control. Problem solved. Thanks>

Ash: Can you please message me in the chat and I'll try to help. I suspect you have your intercept specified differently. Check that you set the factor levels on the Status column so that the virgin status is first.

76. <Nicoletta><Are there any conditions where p should be used instead of padj?>

Ash: Yes, if you wish to rank all of your genes you can use p-value. It is linearly related to padj so the order is the same, but as we don't have NA we are able to rank all of the genes. We will use this later today for GSEA. You should not ever use it for a cut off for selecting a set of "significant" genes, you must apply the multiple testing correction.

77. <Natalie van Dis><To get the top 100 genes by adjusted p-value, would you not rather select by biggest effect size (Log2Fold) in the <0.05 padj set rather than how significant the gene is? Would that not be more meaningful?>

Ash: It is a different biological interpretation. Are you sure that the most meaningful response is the highest fold change? Many regulatory genes will only change a small amount but will have large effects on transcriptional regulation. What we are looking at there are the gene expression changes that are most conserved across our samples, which indicates that they are generally applicable biologically. It's by no means wrong for you to look at the most differentially regulated gene, neither way is "best".

<Natalie van Dis><Ok that makes sense! So in this case, the lower (oops) the padjust value, the more consistent the effect is in all of your samples?>

Ash: Yes, that's correct. Up to a point. You will start to see padj values that are very low, in the 10^{-5} range or lower. To consider a gene with $padj = 0.000001$ to be more significant than one with $padj = 0.0000015$, is putting too much faith in the estimation from the modelling.

78. <Lorea><I'm a bit confused about the number of genes we uploaded in the countdata matrix that Abbi used for the DEseq2 analysis. Is this the shortened one, with 25k genes or the original one with 55k? I understood yesterday that we do not have to filter for lowly expressed genes in DESeq, that it does itself, but the countdata we uploaded this morning was the 25k, was not it?>

<Chandu>Yes you are right, in DESeq2 you need not to do hard filters like what we did in this course. The reason why we filtered is that after filtering we have about 25,000 genes, when compared to 55,000. This makes processing and plotting easy for this course. In real life you may not filter the genes.<Answer>

79. <Name><If the adjusted p value is greater than 0.05 (in all of the genes) but the p value is less than 0.05, can we still use the data but with a bit of a pinch of salt? The fold change is greater than the set threshold of 1.5x compared with baseline.>

Ash: It is vitally important that you pay attention to the adjusted p-values. If all of your padj are greater than 0.05, you can use a higher cutoff, say 0.1 but remember that this will mean you now have a 10% false discovery rate - i.e. 10% of the genes in your "significant" list are false positives. However, there alternative methods of analysis that consider gene ranking that may be more appropriate. We will look at GSEA this afternoon, which doesn't require you to identify a significant gene set, but looks at the overall ranking of the genes. The "pinch of salt" is still necessary even then.

You should also check the p-value histogram for systematic problems with the data, if all the adjusted p-values are high, then it may be that there is some reason why your data does not meet the assumptions of the statistical tests. We will look at this in the next session.

80. <Elias><For question 1.b, are we only interested in significantly upregulated genes, and if so how do you sort those from significantly differentially expressed ones (i.e. excluding down regulated ones)?>

Ash: You'll need to filter the table based on the log2foldchange. Upregulated genes have a positive log2foldchange.

81. <Natalie van Dis><I'm still a little confused about whether we should worry about duplicates or not, and how to handle them. I generated PE data, so from the answers above I understand that you could more reliably exclude duplicates at the gene counting step in the shell. How would you do this? And how to decide if this is necessary or not?>

Ash: You would do this when you run featureCounts, there is an option (check the help page or the manual online) `--ignoreDup`. Whether it's necessary or not is an on-going debate, but for paired end I think it's the better option. The issue is that if we leave them in we are counting some PCR duplicates, which will add error, on the other hand if we exclude them we are losing some genuine duplicates, which will add error. So we need to trade these errors off against each other - which is worse/better? For paired end the number of genuine duplicates is probably significantly smaller than the PCR duplicate rate, so it's probably best to exclude them. But, this would not be true if you are particularly interested in genes that have very short transcripts, especially the smaller non-coding RNAs.

<Natalie van Dis><Ok thanks for the additional explanation. I guess for a Poly-A selection library the smaller non-coding RNAs would already not be included a lot anymore? And the duplicates that you would exclude with featureCounts, are these duplicates in the sense that they all map to the same region or are these the ones that were already noted as duplicates at the raw QC step?>

Ash: The duplicates we want to exclude are based on alignment position - we may have a read that has a PCR duplicate but due to sequencing error the exact base sequence may not match. Yesterday we ran the Picard tool markDuplicates, which generated a new bam file for us. The tool has done what it says, it has marked the reads that are duplicates in the new bam file (we called it "...mkdup.bam"). You would need to use this one to exclude the duplicate reads.

82. <Anika><You said that the more complex interaction model might lead to a loss of power. Is it cheating then, to use batch correction first regressing out the covariate not of interest and then using a simple comparison? What would be recommended?>

Ash: The best way to do batch correction is to include the covariate in the model, you should not adjust the count matrix before using DESeq2 or EdgeR. One way or the other you are going to lose information, so it's better to use the correct model on the raw data.

83. <Anika><Vaguely related. I'm currently working with MicroArray data and have batch-corrected with ComBat in advance, because I'm performing DE analysis as well as other analysis. Would it be better to perform DE on the not

batch-corrected data with the covariate in the model, or is it better to know that all analysis was performed on the same dataset?

Thanks for your previous answer, that's good to know!

Ash: I am guessing you are using Limma for your microarray analysis (yes)? Microarray data is a different ball game because it is intensities rather than counts. Using ComBat is absolutely fine (pfew!!;). I should say that in extreme situations you can do this sort of batch correction with RNAseq data - there is even an implementation of ComBat for RNAseq data see the RUV package. However, this is really for when you have a batch effect but don't know what the covariate causing the batch effect is and so cannot add it to the model.

Anika: Thanks!! **Good to know that the current workflow is ok and how to work with RNASeq if the same problem comes up**

Ash: I should also note that although we are using DESeq2 and have mentioned edgeR, there is an implementation of limma for RNAseq data called voom. It basically transforms the RNAseq data from counts to something that looks like intensities and then runs the limma workflow. It's a perfectly fine alternative, as mentioned in the Blog post by Mike Love that Abbi showed you - the link is on the website.

84. **<Natalie van Dis><For non-model species, is there a way to use the gene annotation file used in the shell directly to add the annotation to DESeq2 results? >**

Ash: Yes, there are a number of ways to do this. Have a look at the AnnotationDBI package. You can for example turn the GTF file into an R database package. In the extended materials for this session there is an example of how to do this.

<Natalie van Dis><Ok great! What is the path for the extended materials on github? I have trouble navigating it.>

Ash: There is link towards the bottom of the main web page:
https://bioinformatics-core-shared-training.github.io/RNAseq_November_2020_remote/Extended_index.html

85. **<James><I'm not sure I understand how the likelihood ratio test is defining which model is better. What is actually being compared between the different models to determine the p-value? How do we know that a difference indicates that it's 'better' rather than 'different'?**

Ash: Thanks for the question. The explanation is quite detailed and I won't be able to cover it fully here, so I'd recommend reading the manuals and looking up the LRT on the web. In simplest terms, the LRT generates its p-value by comparing the deviance in the more complex model to the deviance in the simpler model. The p-value essentially represents the probability that a significant portion of the variance for the gene is explained by the factor(s) that has been removed in the simpler model. The null hypothesis is that the factor does not explain a significant amount of the variance, so if the p-value or rather adjusted p-value falls under our threshold for significance, then we interpret this as meaning the the factor removed does explain a

significant portion of the variance and is therefore necessary to have good model of the gene expression across our sample groups.

<James> - thank you!

86. <Melinda>< I get the following error when running the last annot function:

Error in filter(., n > 1) : ensemblDb::filter requires an 'EnsDb' object as input. To call the filter function from the stats or dplyr package use stats::filter and dplyr::filter instead.

This is my code: `annot %>%`

```
  add_count(GENEID) %>%
```

```
  filter (n > 1) %>%
```

```
  arrange (desc (n))
```

Ash: For reference. This is because the package `EnsDb.Mmusculus.v79` loaded the `ensemldb` package which contains a function ``filter``. We actually want to use the `dplyr` (from `tidyverse`) version of ``filter``, but because the `EnsDb.Mmusculus.v79` package was loaded last, its ``filter`` is the default (you will see messages about "masking" during package loading). You can get around this by explicitly saying which package you want to use like this: ``dplyr::filter``. However, as we are going to use `filter` a lot, it is probably easier to restart R and be sure the load the `tidyverse` package last.

<Tutor><Answer>

87. <Munise><How do you go from the annotation table we made to the pre-annotated table we used? Does that require manual work?>

Ash: Yes, you need to make some decisions about duplicated entries. One way is to spend some time with Ensembl and Entrez on the web and try to figure out which Entrez ID's you want to keep. This can be time consuming and is not always worth the effort as often these duplicates entries are for genes that are not of interest. It's a judgement call that you'll have to make. In the extended materials there is a section on annotating via `biomaRt`, in their we mention one helpful method, which is to pull the gene symbol that the NCBI have annotated the Entrez ID with. Often only one of the duplicates will have the same symbol as is in the Ensembl annotation, this will usually deal with most of the duplicates.

88. <Munise><Is shrinking something we have to do or does it depend on the spread of fold changes?>

<Abbi><We use shrinkage for visualisation because genes with low counts and high FC will appear to be much more significant on plots than they are if we don't. `lfcShrink` just shrinks the log fold changes towards zero to compensate. It is just a way of getting better visualisation and ranking of genes.>

89. <JAgon><What does this mean "Error: Join columns must be present in data. Problem with `GENEID`. Got it! thanks

90. This could mean you have spely the column name GENEID wrong, or that this column is not present in both of the data

91. <Eve> I get this error whenever I try to do the rename: `Error in rename(., logFC = log2FoldChange, FDR = padj) : object 'log2FoldChange' not found`

<Abbi><R is looking at the rename function from a different package. This is the masking we were talking about with functions having the same names in different packages. Whichever package was loaded most recently supersedes the previous. You can either type `dplyr::rename` to explicitly tell it that you want that version of the rename function or restart R and this time make sure tidyverse is the last library you load.>

<Eve> Got it, thank you!

92. <Lorea><R scripts versus R notebooks: is there any reason why this afternoon we are using a notebook rather than a script? I'm asking because I found notebooks more confusing and I don't understand if this is a personal option but the exercise could be done in an R script too. Thanks>

Ash: Yes, you can do it either with R script or R notebooks. When you do come to do analyses yourself R notebooks are a better way to record your process as you can write text between the chunks. In addition you can render them straight to pdf or html reports by clicking the "Preview" button. All of the web pages for this course are made with R notebooks. The other advantage is that the output is in the same document as the code, so you can just read down through it to see what you have been doing and what the results are.

93. <Lorea><Regarding the curated annotation we are using, I am not entirely sure where we can get it for our own experiments. Is this the same as the .gtf file we used in day 1? >

<Abbi><The curated annotation came from Ash painstakingly fixing all the issues Chandu illustrated could happen with the miss and multi matching so that we didn't have to worry about that today. You will have to do it yourself for your organism. There are some pointers in Ash's answer to question 87.>

94. <Natalie van Dis><Could you maybe explain a bit about how the clustering is done by the heatmap function? And could you extract those clusters and examine them? Like what type of genes cluster together? Also how to decide in how many groups to split them for the final plot?>

Ash: The clustering is a hierarchical clustering based on Euclidean distance between the genes. You can achieve the same clustering using the `hclust` stats command. The help page of that function is probably the best place to get an idea of how it works. You can extract the clusters, if you assign the Heatmap to an object `myHM <- Heatmap(.....)` rather than just printing it, then the object includes the information about which genes are in which clusters. I can't remember exactly how to access them, you'll need to have a look at the ComplexHeatmap help pages/manual.

95. <Natalie van Dis><Ok thanks for the explanation! A related question: Could you decide to do clustering first and then do DE analysis with the clusters rather than the genes? And if that's possible, when would you want to use this approach?>

<Chandu> Gene set enrichment dose similar thing, what gene sets show significant enrichment between two conditions.

96. <James><just wondered if there was a way to change the colours of the cell type and status annotations? The automatic ones are ugly!>

<Abbi><Yes! The colours are chosen randomly each time you run it unless you set them and more often than not you get horrible pairings. The documentation for this is here

<https://jokergoo.github.io/ComplexHeatmap-reference/book/heatmap-annotations.html> and you just specify the colours in the heatmap annotation section. You are looking for col argument in HeatmapAnnotation if I remember correctly, you change it in the bit where we make ha1 in the html code. There is a huge amount more you can do with this package, some super cool plots!>

97. <Anika><How often are gene sets updated? I've had to re-run an analysis from 9 month ago, and now found that COVID-related genesets are enriched (in my liver injury genes..). I assume this happened because the geneset list was updated, but how do I know which version clusterprofiler is using? >

<Chandu>Difficult to generalise, depends on which database talking about. For example MiSigDB updates every year or more.

98. <Lorea><Following on Ash's comment, I cannot get the slides in the webpage. When I click on slides, I just see the first introductory slide, but not the ones used for the presentation. I tried in Github but I couldn't find in which exact folder they are. Thank you>

<Abbi><They are working fine for me, I'm in chrome, what browser are you using? Also try refreshing the webpage, sometimes the links need to update if things have changed. On github Ash's slides for this are in the html folder and are called 06_Introduction_to_Functional_Analysis_in_R.html>

<Lorea> < I can see there are two types of files, some you can download easily, such as when I click on this link from day 2: Experimental Design of Bulk RNAseq studies - Abbi Edwards, and others where I see the presentation in the browser but I cannot figure out how to download the slides, such as the 'Introduction to RNASeq Methods' from Jon Price. Maybe this is the way it works, I just want to make sure it needs to be seen on the web. In any case, I think you mentioned that the link to the course online will be available if we want to check something in the future?

<Abbi><I've checked safari too and it works fine for me there. Yes the course website will be up as long as github exists, plus the recordings of us giving the course will be up for the next month and as long as you download them to your machine before the month is up you will have them for as long as you want. If you

want to download particular slide sets the easiest thing is to get them from the github site by cloning the repo or downloading it all as a zip file.

99. <Leila ><how the Gene enrichment score is calculated?>

<Chandu> Nice explanation here :

<https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html> Look in 'GSEA Statistics' section.

100. <Leila><How we can get ride of the sex linked genes in our gne list before starting the analysis(even before getting the pvalue and padj or even normalisation.)>

Ash: If you wish to remove specific genes you can simply find their IDs and then filter the count matrix to remove them. This is not recommended in most cases. For example if you are concerned that sex bias in you sample groups is causing sex linked genes to appear as differentially expressed - e.g. you have control and treatment, 6 reps each, but 4 females in one group and 4 males in the other - then you should include "Sex" in the model (~ Treatment + Sex) in order to take account of this. If you have biased you groups too strongly for the modelling to take account of the bias (e.g. all males in one group and all females in another) then your experiment is fundamentally flawed. You can remove the definitively sex linked genes, but then are you sure that changes in the other genes are not a results of differences between males and females.

101. <Rob H><I've been getting an error with the following piece of code but I can't spot what's wrong:>

```

1 library(clusterProfiler)
2 library(tidyverse)
3 load('Robjects/Annotated_Results_LvV.RData')
4 library(pathview)
5 library(org.Dm.eg.db)
6
7 head(shrinkLvV)
8
9 sigGenes <- shrinkLvV %>%
10   drop_na(FDR, GeneID) %>%
11   filter(FDR < 0.01 & abs(logFC) > 1) %>%
12   pull(GeneID)
13
14 backgroundGenes <- shrinkLvV %>%
15   drop_na(FDR, GeneID) %>%
16   pull(GeneID)
17
18 ego <- enrichGO(gene=sigGenes,
19                OrgDb = org.Dm.eg.db,
20                keyType = 'ENSEMBL',
21                universe = backgroundGenes,
22                ont = 'MF',
23                readable = TRUE)
24
25
26

```

16:15 (Top Level) ↕

Console Terminal × Jobs ×

~/Course_Materials/ ↕

```

· ego <- enrichGO(gene=sigGenes,
·                 OrgDb = org.Dm.eg.db,
·                 keyType = 'ENSEMBL',
·                 universe = backgroundGenes,
·                 ont = 'MF',
·                 readable = TRUE)
·
· -> No gene can be mapped....
· -> Expected input gene ID: FBgn0037347,FBgn0037108,FBgn0037926,FBgn0036569,FBgn0027080,FBgn0030932
· -> return NULL...
·
· |

```

Ash: Do not drop any genes from the backgroundGenes, we want all the genes in our dataset. You have specified org.Dm.eg.db, which is the drosphilia database, but our data is mouse, so the gene ids do not match.

102. <Leila ><If we have multiple separate bam files how to integrate them into one bam file with all the data from all the samples combined? >

Ash: You can merge multiple bam files using samtools or Picard, see the help pages. On the other hand, generally we would keep samples in separate bam files.