Analysis of bulk RNA-seq data

19th - 21st May 2020
9:30am - 17:30pm (Hopefully not quite that late!)

Links

All the materials for the course can be found at the **Course Website**

- Please connect to the course via this Zoom link: https://us02web.zoom.us/j/83902459004?pwd=K005TVJBemZPVGluWUQxVkRISDNKZz09
- Course feedback survey: https://www.surveymonkey.co.uk/r/WYNVP97
- **Course participant introductions**: https://docs.google.com/document/d/1Q5_E4vjVcRkYaHRc-D4EQ4EBCWzShjPPvK3qulCZE9U/edit?usp=sharing

Course Video Recordings

Each days lessons will be recorded and linked here shortly after the day's teaching:

- **Day 1 :** https://us02web.zoom.us/rec/share/xuFVBbro5lJOQ53K413bYqAZO6m1T6a81HRIrvYMnUf4kRmD5NCcCtDHRZMaZhZ- **Password: 5J*?6R%n**
- **Day 2 :** https://us02web.zoom.us/rec/share/19V6coju-lxJS7OT-XnAa5M_DI7Meaa8gHAZ_KELyhyY4DynX6wsPelSQy3BeqSj **Password: 4U@7&2$+**
- **Day 3 :** https://us02web.zoom.us/rec/share/xO5fJfLf2j1IZ6_h5XvxaKMxJZn-aaa81yQcq_AJn057un05KDSXrGcWiDOBBatp **Password: 9n=Lm4%%**

Remote Machines

During the course we will use two remote servers to run both command line tools and R. You should have received an email with links, a username and a password.

Sign up to our mailing list to get notifications of upcoming courses from the Bioinformatics Training Facility

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 Questions

1. **Holly Craven I'm having some slow connection speed issues - can i have the menti code again please?**
   <Abbi> Don't worry, that was just for Ash to check who had used the command line before, each time there is a new menti quiz we will give the menti instructions again.

2. **Gabriel Rinaldi: can we use our local R studio software?**
   <Jon> Sorry, there may be a problem accessing the data on your local machine, is it possible for you to use the virtual machine?.
   <Gabriel Rinaldi> yes, I am using the virtual machine OK

3. **Nermina Lamadema Why would I want to use polyA over the other types of library prep?**
   <Jon>This really depends on the biological question you are asking. The two main types are Total RNA and Poly-A selection. Poly-A enriches for mature mRNA but does not include some species of RNA within the cell that lack a poly-A tail.

4. **Shaimaa Hassan: how to decide the coverage?**
   Zeynep: It depends on the research question and the experimental design (which is covered on slide 5). Briefly, for in-depth analysis you will be aiming at ~100M (or more) reads and for a general view 5-25M reads would be enough.

5. **<Gabriel Rinaldi> can we download and store locally these powerpoint presentations?**
   Zeynep: You can download and store them.

<Gabriel Rinaldi> how can I download the presentation? I can't find the option…
Abbi: Everything is stored on the github site (link at top of webpage)
Paul: The slides are in the html folder, if you download the git repository you should be able to open them with your browser if you need a local copy.

6. **<Gabriel Rinaldi> I guess the concept of the biological replicate depends on the experimental unit and the biological question behind the study (?)**
   <Jon> Yes certainly, this will depend heavily on the organism and experimental design as to what exactly you will call a "replicate". If you are unsure when planning an experiment it is best to talk to a bioinformatician before starting.

7. **Nermina Lamadema How many replicates are considered sufficientMine is already been covered off-line thanks**

8. **Q: Will we cover power analysis in this course?**
   Dom: No. The topic will be briefly mentioned during the stats part of the class on wednesday afternoon.

9. **Esther Palomino Why do you know that it is a bad result?what mean each element of the graphic?I was asking for the first graphic, what mean the yellow bar. :)>**
   Ash: The graph is showing the range of base quality scores at each position across all the reads. We would like high base quality scores - 35-40 - to be happy that the base call we have is reliable. If the scores is low, we can trust the call - what we have as an A may be a C. In the case of the bad plot we can see lots of low quality calls. In truth, we very rarely see this anymore, the sequencers are much more reliable, this bad data we are showing is very old.

10. **Nermina Lamadema Should you not check your RNA quality and library prep quality priort to seq run?**
    A: Yes, absolutely, but that still doesn't guarantee sequencing quality, or that the cDNA is what you think it is. QC is vital at every stage.

11. **Shaimaa HAssan from where you get the adaptor contaminant>**
    <Jon> Adaptors are present during library preparation to allow the RNA sequences to bind the flow cell. They are present in all sequencing experiments and are subsequently removed. You may find they are present in sequencing data you recieve (You can find out with FastQC). If they are present then you can use a trimming program (e.g Trimmomatic) to remove them.
    Sankari: This kind of contamination happens when the DNA content is low and adapters dominate that you would get the empty adapters getting sequenced more than DNA. This leads to loss of your data. However, current library preparation kits have developed the concentration of input sample and adapter

content better so that we don't get adapter contamination much. However, if the input is low, you would see them.

12. **Holly Craven Would the 'bad data' rule still apply if you had a genome with a skewed content, eg plasmodium, which is around 80% AT rich?**
    <Ash> No, for the GC content you obviously have to consider what GC content you are expecting. The bimodal distribution that Abbi showed was actually due to contamination of Drosophila sample with RNA from bacteria (the plates were infected).

13. **Ren Moon I've had the hand raised for a while but I can't access the guacamole. The link and sign in works, but it keeps saying connection is lost because the server is taking too long to respond. Is there another way I can do this**
    A:If you are having technical issues, please send Paul Judge a personal chat. He will try to help.
    <Abbi> If you are getting this with Safari, try a different browser
    <Paul> The main things that cause issues with the guacamole login are particular versions of Safari (We recommend switching to google chrome as that's what we tested with) or adblock/noscript plugins. If you are having issues try to turn off extensions and connect with chrome.

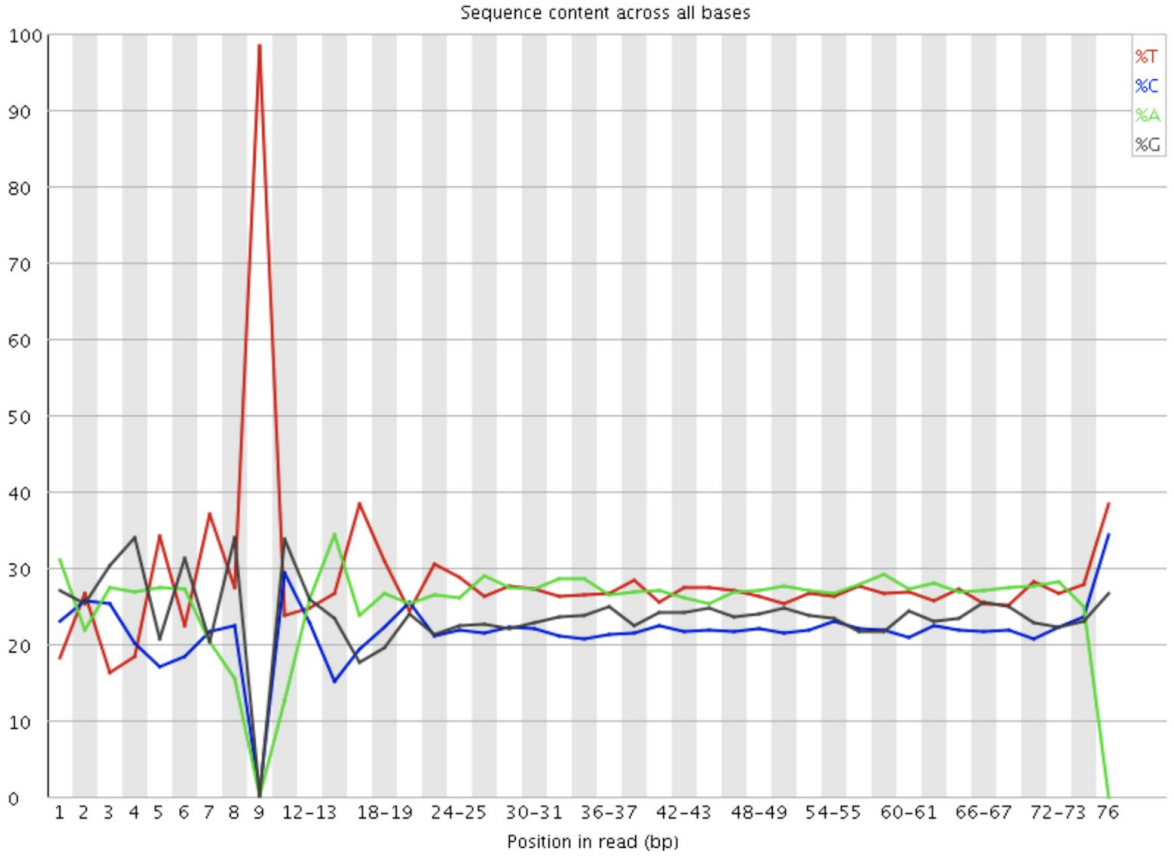
14. **Fernando If I would like to run Fastqc on my computer, what should I do?**
    <Paul> You can download fastqc at the project page here:
    https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
    It's java based so should run on most operating systems without too much trouble as long as you have a working java installation. You might find it easier to work in the environment provided today but I'm happy to help with this and other installations as the course progresses.

15. **Georgia Does the read depth affect the per base sequence quality at all, for example if I have data that was just a very small spike into someone else's run would it be reasonable to expect this to look a bit worse due to random noise?>**
    <Ash> The base quality is independent (largely) for each read. In the case you describe, contamination with reads from some other library, it would depend on the base quality of that library. If the sequencing for that library is good, then you would not notice anything in the per base quality.

16. **Shaimaa I always couldn't understand what the Per tile sequence quality?>**
    <Ash> The per tile sequence quality is looking at the overall read quality at physical locations on the flowcell. The flowcell is sort of like a slide, so that plot is looking down on the slide and highlighting regions of good/bad quality. In the past it was common to have problems such as air bubble causing a region of poor quality sequencing and you'd be able to see that here. In reality this sort of thing is rarely a problem these days.

17. **ther Palomino on the terminal I can not write "~" in my laptop. It is spanish and I can not find that option in Hi I'm a bit stuck at step 6 do you think you could help please?**
A:Could you raise your hand in zoom and one of us can chat with you. (It is on a spanish keyboard somewhere as I have used on in the past, I just can't remember how to get it off the top of my head).
Zeynep: Alternatively you can write `/home/ubuntu` instead of ~
I thought I had to change it to participant

18. **Shaimaa Hassan Regarding the per base content f I got something like this what does it means**



<Ash> Bases 24 onwards look fine. The initial bases always have this pattern if you used random primers in the RT as there is some bias in the production of the primers and the PCR. However, the strange spike in T's at base 9 is new to me. This might be down to something in the library prep protocol - what was it?
NEXTFLEX Rapid Directional RNA-Seq Kit
Sankari: Maybe it's from primers used in the library preparation? Or the initial part is not trimmed?
Shaimaa: I had a very small read number and also contamination from other source is this may be the reason why?>>

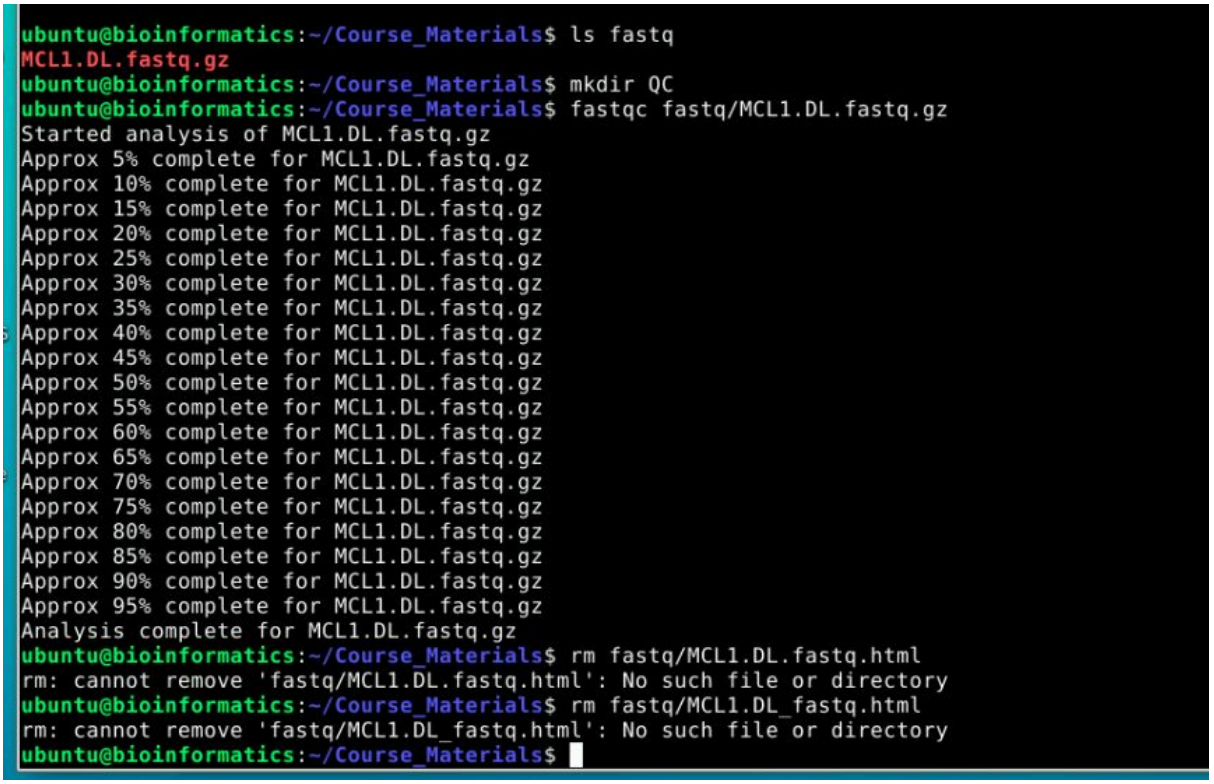
19. **Fernando How I can use the -o in the question 6, I cannot go futher>**

<ash> `fastqc -o QC fastqc/MCL.DL.fastq.gz`

20. Jenni Karttunen How do I open the html report in a browser?>
    <Abbi>Click on the folder symbol is the centre at the bottom of your screen, this will bring up your folders, if you navigate to your html report and double click on it, it will open in a browser. Also typing firefox and the html on the cmd line will work.
    Zeynep: If you have been working on the ssh you need to switch to desktop to see the report.

21. Komal I do not understand the sequence duplication levels and deduplication for the reads.>
    A: The sequence duplication level refers to how many times we are seeing identical reads in the data. In general we don't worry too much about this at this stage as we will look at it after alignment. The FASTQC report is designed for use with all sorts of sequencing data and in this case this plot is not too useful for us yet.

22. <Gabriel Rinaldi> Sorry I am a bit slow, I am following the exercises and when I used the rm command to move the report to the QC directory I am getting this error:



A: Use `ls fastq` to see the files in the fastqc directory and check the name of the file that you want to delete. If you want to chat in a breakout room with a trainer please raise your hand in Zoom

23. **Shaimaa Hassan If I work with published data that has SRA number how can I check its quality by fastqc>>**
    <Abbi>You will need to download the fastq files and then run fastqc as we have here
    Ash: When you download the data it may come as 'sra' files, this a special compressed format they use for storage and transfer of data. SRA has a suite of tools that you can use to transform to fastq.

24. **Fernando I cannot run fastqc -o QC fastqc/MCL.DL.fastq.gz for the number 6**
    <Jon> What is the error you recieve? Did you create the QC directory?
    cd ~/Course_Materials
    mkdir QC
    fastqc -o QC fastqc/MCL.DL.fastq.gz
    <Fernando> Thank you I see what I was wrong

25. **Maria Do I need to use Linux when I come to analysing my data?>**
    Zeynep: For certain steps (such as alignment and quantification) you will need Linux, but once you start working with counts (ie. gene expression data) you will be using R, which can be used through Windows as well.

26. **Shaimss whats different between salmon and hisat2?**
    <Abbi> These are two different pieces of software that both do alignment, salmon is a new class of aligner called a pseudo aligner, although fast, the trade-off is that you don;t receive bam files for investigation. They use different methods for doing the mapping. Salmon uses paired end data but HISAT2 can use both single and paired end. Your choice of aligner will depend on your experiment and what you are looking to find out from your data. Different aligners have pros and cons.

27. **Fernando Now We run an example, but how can I do If a have one file, should I install FastQc in my computer?>**
    Zeynep: You can download fastqc at the project page here:
    https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

28. **<Gabriel Rinaldi> How important is to DNAse the RNA for RNAseq? Given potential DNA contamination in the sample, and hence reads coming from introns? I know some researchers don't bother to DNAse the samples.**
    Sankari: We recommend using DNase in your RNA preparation. Especially for RNA which are non-coding or close to have less introns. For mRNA samples, it's more of a problem during qPCR experiments if you use random primers. Most of the mRNA library preparation kits deal with oligodT primers. If mRNA is converted to cDNA using oligodT primers, your DNA contamination will be reduced. Generally having DNase wont hurt and the same sample can be used for both qPCR and sequencing.

29. <Justina><What is the difference between STAR and HISAT2? Which one would you recommend using when? There is also Bowtie2, then would I use this?>
    <Abbi>In reality there will be very little difference between using STAR or HISAT2. They do the mapping slightly differently and HISAT2 is a bit faster but essentially using one or the other won't affect your biological outcomes at the end. Bowtie is using the tophat aligner, this is quite outdated now and I would suggest STAR or HISAT2 instead.
    <Justina> Just to follow up on this, Bowtie2 is also outdated?
    <Abbi> Yes sorry, tophat2/Bowtie2 has now been replaced by HISAT2

30. Jenni Karttunen><Which alignment tool do you recommend for small RNA seq data? miRNA?>
    <Jon> Which sRNAs do you plan to investigate? SPORTS is good if you don;t know what you are interested in. Then for each species of sRNA there are specific tools for quantification. sRNABench and miRDeep2 perform well for miRNA.

31. Shaimaa Which aligner best for lncrna novel detection and not well annotated genome?
    <Abbi>This is a bit of a tricky one, I have used HISAT2 for this in the past with Trinity for the novel gene/lncRNA detection
    (Shaimaa) Whats trinity?

32. Justina I was trying to copy paste the command line but it did not work. Does not cm+c cm+v shortcuts work for the terminal? I have a feeling they used to work.>
    A: Because it is a remote machine you need to use Cmd + Shift + V, I think, but actually I've not been able to get it to work either. It would work in a local terminal on your machine. Tomorrow we will be using RStudio and it works fine.
    Paul: You can also use the shortcut "Ctrl+Alt+Shift" or "Ctrl+Cmd+Shift" to open the side panel and access the clipboard directly

33. Shaimaa im not sure I quite understand this hisat2_index_base write ht2 data to files with this dir/basename I don't find this directory and why we didn't add -f option as I think it's fasta file not fastaq file and according to what we determine the number of thread to use some times p7 other p8

34. Komal><Does indexing the genome basically divide it into parts, and then the query seqs are searched in each index?>
    Zeynep: To facilitate rapid lookup, the genome is hashed or transformed, and then indexed. HISAT2 used Burrow-Wheeler transform (BWT) and Ferragina

Manzini (FM) index. This paper provides a nice primer to these approaches: https://www.ncbi.nlm.nih.gov/pubmed/19430453

35. Alyce How do I set the number of threads to 7? Can't find it in the help section>
Zeynep: -p 7 (when using hisat2) - but it will be different for different tools. You would find these options in the manual of the tool (or when you type <command-name> --help)

36. Komal <As HISAT2 is used for aligning RNAseq reads, which tools could be used to align short DNA reads to the genomic regions?>
Bowtie2 can be used for short DNA and is suitable for things like ChIPseq when are only interested in location and not exact mapping of each base. If you are doing variant analysis then you need a more precise mapping and bwa is the one that we generally use.

37. Shaimaa I got this error

```
ubuntu@bioinformatics:~/Course_Materials$ hisat2 -x references/hisat2_index/grcm
38. -U fastq/MCL1.DL.fastq.gz -s shaimaa.sam
Could not locate a HISAT2 index corresponding to basename "references/hisat2_inc
ex/grcm38."
Error: Encountered internal HISAT2 exception (#1)
Command: /applications/hisat2/hisat2/hisat2-align-s --wrapper basic-0 -x referer
ces/hisat2_index/grcm38. -s shaimaa.sam -U /tmp/1495.unp
(ERR): hisat2-align exited with value 1
ubuntu@bioinformatics:~/Course_Materials$ xx
```

>also got this error, I do not have a typo as far as I can tell
Apologies, the error I have is "hisat2-align exited with value 1"
<Abbi>It looks like you have a typo in the name of your reference index

```
ubuntu@bioinformatics:~/Course_Materials$ hisat2 -x references/hisat2_index/mmu.GRCm38. -U fastq/MCL1.DL.fas
-s bam/shaimaa.sam -p7 -t
Could not locate a HISAT2 index corresponding to basename "references/hisat2_index/mmu.GRCm38."
Overall time: 00:00:00
Error: Encountered internal HISAT2 exception (#1)
Command: /applications/hisat2/hisat2/hisat2-align-s --wrapper basic-0 -x references/hisat2_index/mmu.GRCm38.
m/shaimaa.sam -p7 -t -U /tmp/1567.unp
(ERR): hisat2-align exited with value 1
ubuntu@bioinformatics:~/Course_Materials$
```

no error I use tab so no way there is typo
Zeynep: the error is "Could not locate a HISAT2 index corresponding to basename "references/hisat2_index/mmu.GRCm38."
You need to remove the dot at the end of the index file.

38. Georgia I am a bit confused about the terminology in the description of the SAM format terms. What exactly do they mean by template? And are the RNEXT and PNEXT terms only relevant to paired end reads?
<Jon> RNEXT and PNEXT are indeed used for paired end data and they tell you where the paired read has been aligned.
You can use template and genome reference interchangeably

39. **Michela How can you run your alignment "in real life" on your computer with real data? Do you need a Unix interface or is there any less complex way to do it?>**
<Jon> I would always recommend getting comfortable with command line. If on a mac you can install of these tools. With windows it is a bit harder.
Your institute may have computational resources you can use (most have servers for doing this kind of work). As a last resort there are online "easier" tools to use such a galaxy but these can be less flexible.
<Abbi> Some aligners will also need more computing resources than are available on your personal computer to process in a reasonable amount of time so I would suggest investigating what is available at your institute as Jon says.

40. **Jenni Karttunen My exercise 2 alignment is still running, probably did something wrong… how can I stop it? OH, now it finished maybe after 30 min!>**
Zeynep: You can try pressing ctrl and C simultaneously.

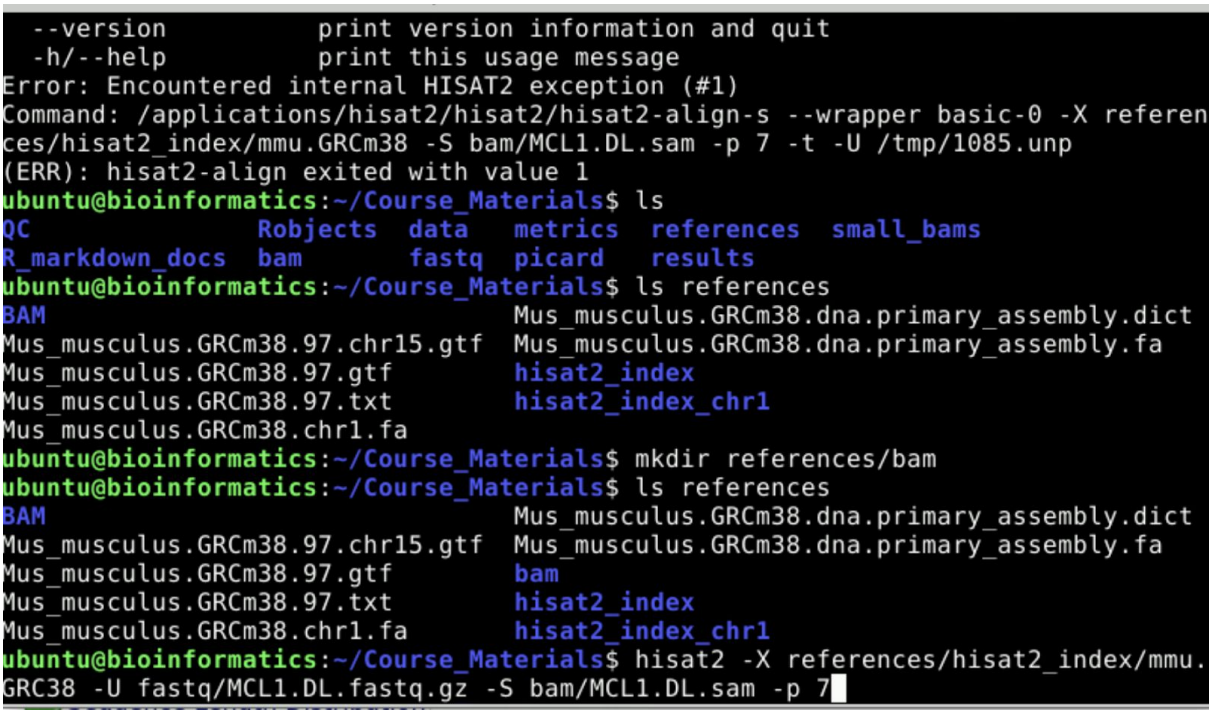
41. **Fernando Could we have the screenshot of the last steps in exercise 1 ,2 and 3 ? I think that I am doing something wrong when I type. Or maybe a text file to try to repeat the exercise>**
<Abbi>We'll make all the solutions available on the webpage at the end of the day :)

42. **Ren I have checked what I typed and can't see any immediately obvious errors but keep getting ERR: "hisat2-align exited with value 1", encountered internal hisat exception**

```
  --omit-sec-seq     put '*' in SEQ and QUAL fields for secondary alignments.

 Performance:
  -o/--offrate <int> override offrate of index; must be >= index's offrate
  -p/--threads <int> number of alignment threads to launch (1)
  --reorder          force SAM output order to match order of input reads
  --mm               use memory-mapped I/O for index; many 'hisat2's can share

 Other:
  --qc-filter        filter out reads that are bad according to QSEQ filter
  --seed <int>       seed for random number generator (0)
  --non-deterministic seed rand. gen. arbitrarily instead of using read attribut
es
  --remove-chrname   remove 'chr' from reference names in alignment
  --add-chrname      add 'chr' to reference names in alignment
  --version          print version information and quit
  -h/--help          print this usage message
Error: Encountered internal HISAT2 exception (#1)
Command: /applications/hisat2/hisat2/hisat2-align-s --wrapper basic-0 -x referen
ces/hisat2_index/mmu.GRCm38 -u fastq/MCL1.DL.fastq.gz -s bam/MCL1.DL.sam -p 7 -t

(ERR): hisat2-align exited with value 1
ubuntu@bioinformatics:~/Course_Materials$ hisat2 -x references/hisat2_index/mmu.
GRCm38  -u fastq/MCL1.DL.fastq.gz -s bam/MCL1.DL.sam -p 7 -t
```

<Jon> Could you send your command please?
Zeynep: When you specify the output sam file, you need to use -S instead -s

Ren: this gives me the same error message
Jon: -u is also a capital -U
Ren - I've changed that one too, and the X. Sorry
Jon: still doesn't work?
Ren - No
A: is this how your command looks now?:
hisat2 -x references/hisat2_index/mmu.GRCm38 -U fastq/MCL1.DL.fastq.gz -S bam/MCL1.DL.sam -p 7
Ren: Yes. Still gives me the same error
(**Kyungtae:** I had the same issue.)
A: Could you please check 1. The fastq file has the correct name. 2. The reference also has the correct name and is present in the right directory and 3. The bam directory exists.
Then could again screen shot the whole output please.



Solved: Typo in the reference genome name.

43. **Benjamin** I'm working on a macbook pro and I can't find the shortcut for the pipe symbol. I usually used alt+shift+l on my mac but it doesn't work on the online terminal. Could you tell me what I should type?>
A: Pipe symbol on a mac should be shift + \
(near the enter)

44. **Shaimaa** Is the trimming something essential before alignment especially If I have good score? For adaptor?>
<Abbi>Not essential, it will depend on your data. If you some of the issues highlighted in the Fastqc section you will need to do it, eg. adaptor contamination. There is an example in the extended material section.

45. **Shaimaa><I wonder which is the best genome file to use soft masked or hardmasked im looking for novel LncRNA>**

And when I align use Hisat2 what you mean by trinity?

A:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3571712/ - Trinity software

46. **Shaimaa And regarding the softmasked and hardmasked genome>**

Ash: HISAT2 converts all lower case bases in soft masked genomes to upper case. I.e. by default it ignores the masking. It then indicates in the read alignments whether the read is mapped multiple times. I am not sure about other gapped aligners. Generally you should use a soft masked genome and then filter your reads later on according to what it is you are looking for.

47. **Alyce I've run the picards tools but how do you read them to find duplication rate? Is in the text file?>**

A: Yes it's in the text file. It's not the easiest to read, but you can find it. At the practical we'll generate a report, which makes life easier.

48. **Romit** How do you look inside the bam/MCL1.DL.alignment_metrics.bam file to find the mismatch rate?>

Ash: The output of the alignment metrics should have been called .txt not .bam, my fault. But it doesn't matter, it is a plain text file you can view it on the command line by typing

`cat bam/MCL1.DL.alignment_metrics.bam`

<romit> yes i tried 'cat' and the output was gibberish.

Ash: It must be the wrong file. If you can't figure it out, raise your hand and someone can help you.

49. **Maria Does caps lock not work in Linux or the Command Line?>**

Jon: Yes it should work.

50. **Shaimaa I cannot run the Collect RnaSeq**

Ash: You have spaces after the `=` signs, take them out.

Shaimaa Thanks it really helped

51. **Shaimaa I don't understand how I know if my library is first stranded or second stranded or unstranded**

Jon: This is determined during the Library preparation step, you can always find this out by talking to the person that generated the data.

The person who did the library or who did the sequencing?

Because I'm preparing the library but sent it to illumina for sequencing

>It should depend on the kit you are using to prep the library.

52. <Justina><Any good alternatives to Picard?>
    <Tutor> <Answer>

53. <romit> some informatics cores produce reports that look for contamination with fungal / bacterial / grass etc. Do we just build additional reference genomes into our pipelines to look for these sources of contamination during library preps?
    <Abbi> We do this at the CRUK CI using software called MGA which is available online. It basically takes a sample of your reads and runs it against multiple genomes to check for this. You can do this too, we don't cover it here but you would run it in the same stage as doing your fastqc reports.
    Ash: Our MGA (multi-genome alignment) tool is available here: https://github.com/crukci-bioinformatics/MGA . It takes a bit of work to get it installed and you would need to build indexes for all of the genomes that you want to test. The advantage with it is that it uses a lighter weight tool - Exonerate - to do the alignment and only aligns the first 36 bases of a random 100,000 reads. This is enough to get an idea of the different species and is much faster than doing a full alignment. It also produces a handy report.

54. <Dawei Sun><Could we remove the PCR duplicate and how? Is this necessary?>
    <Jon> For DNASeq this is a necessary step. But for RNASeq it is not advised.

55. <Shaimaa> What is the secondary read mean is it the reads mapped to different location of what?

A: When a read can map equally well to multiple places in the genome hisat2 will return all of these alignments (up to a limit, by default this is 5) and will choose 1 as the "primary" alignment and mark the others as secondary. You can adjust the number of secondary alignments it returns. Any read with secondary alignments will have a mapping quality less than 60, we generally refer to these as multi-mapped reads.

56. <Georgia> What is the difference between multi-aligning reads and multi-mapping reads (e.g. if you leave in the --primary flag but include the -M flag compared to if you take out the --primary flag but include the -M flag)

    A: If you have only the -M tag it will count all the alignments of a multimapping read. Adding the --primary means that it only counts the primary alignment. I am not sure if anything happens if you only add --primary and not -M as this would not make much sense.

57. **<Dawei Sun > Does the gtf annotation file and the genome file, which we aligned to need to be from the same source like both ensembl or both UCSC? Would it create some confession if not from the same source?**
    Ash: featureCounts will actually cope with the difference between USCS and ensembl chromosome names and will return results even if you are using the annotation with 'chr1' when you reference genome was '1'. Generally it's best to stick to the same source for all of your references, however, it may be that you want to use the UCSC browser but have Ensembl gene annotations.

58. **<shaimaa> I dont quite understand the multimapping?**
    >
    Ash: Multimapping is when a read will align equally well to multiple places in the genome and so we can't actually know for sure where the fragment originated. This may be because the read originates from a repetitive region of the genome, or perhaps in a gene that this part of large family of similar genes, or maybe the read is just poor quality and so there are lots of place that it could be mapped equally "badly".

59. **<Shaimaa> Feature counts is raw number not RPKMs?>**
    A: Yes, raw read counts.

60. **<Dawei> For featureCounts, is there a sort of standard percentage of aligned sequences for judging whether the experiments went well?>**
    A: This will vary considerable depending on the species and the actual experiment so there is no single answer. Really, this is just the same as the RNAseq metrics plot that showed coding/UTR etc.

61. **Q: Would it be possible to spend just a couple of minutes at some point giving a very brief description of what parts are missed out of the abridged online course, and where these are in the course materials in case we find in analysis of our own data that we need to draw on these resources?>**
    A: All the additional materials are linked in the main document under "Extended materials". The only thing we skipped today was read trimming.

62. **<Dawei> Just want to be clear about this, when we only use the --primary in featureCounts does this we are also aligning the multimapping reads using it's primary alignment or we are completely not taking them into account? Simliary**

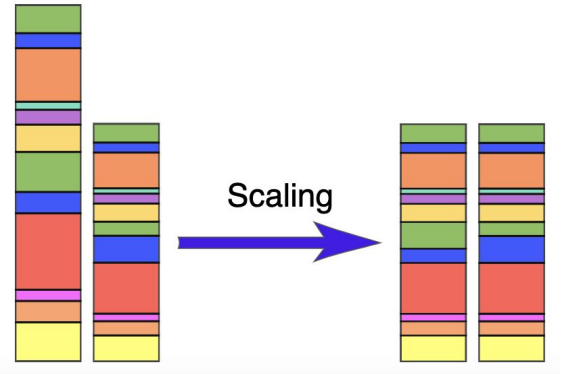
when we are doing -M function, are these multimapping reads aligned multiple times to different locations?>
A: Yes, using -M counts all of the alignments for multimapping reads, adding --primary limits this to just the "primary" alignment.

63. **<Romit> When do we use exon counts in preference to gene counts?>**
A: This may be of interest if you are looking at differential splicing events.

64. **<Shaimaa> regarding the alignment how to trim the adaptor and is it standard that the adaptor length 9 bp?>**
A: There are a number of tools for read trimming. In the Extended materials we give and example using Trimmomatic, you can also look at CutAdapt and Fastp. The adapter length is variable. There should be no adapter at all in your reads under ideal circumstances. The adapter only appears when the original RNA fragment length is shorter than the read length. E.g. if the original fragment is only 80 bases and the read length is 100 bases, then the sequencer will read 20 bases of adapter. In any library the fragment length varies from read to read and depends on the library prep.

Day 2 Questions

65. **Sudipta :In the normalisation scaling example, are we not losing out on the difference between A and B due to the normalisation process?**



<K> The visual depiction you see here is highlighting the difference in sequencing depth. It does not reflect the changes in the expression. We need to make sure that unequal depth does not falsely reflect it.

66. **Sudipta: If high variation is observed across multiple samples in the same set of genes, should we be taking those into account instead of the software ignoring them?**
    <K> You expect genes to show similar expression levels among the replicates of a group. If some genes are showing very high variation within a group, it could mostly be attributed to some kind of technical noise (so should be ignored so that it does not affect the overall analysis). This stands true even for replicates that are expected to be closely related (clustered) within a group. So if you observe outlier samples (based on clustering), they should be removed prior to the analysis.

67. **<Esther><Probabily it is a stupid question but: If we can do with R the same than we did yesterday, Why we did it?...ok, thanks >**
    Ash: Yesterday we were working with the full sequence data, the files are very large and many of the analyses we were doing were very computationally intensive. R does not handle high memory tasks well and we really need to use a more powerful means of doing this. There are ways to do some of things we did yesterday with R, e.g. the featureCounts, but it is much much slower.

68. **<Anna >< The 'logcounts <- log2(countdata, +1)' command gives me an error: Error in log2(countdata, +1) : 2 arguments passed to 'log2' which requires 1. What does this mean?**
    Sankari: Please remove the comma.
    Thank you! Works now
    Sankari: It's an arithmetic operation on the whole data. So no need for commas.

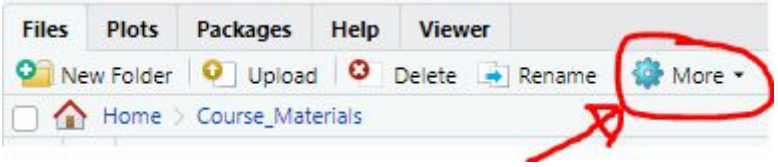
69. **<Gabriel Rinaldi> removing genes: I guess you can remove genes if you have previous information, if not, why should you remove some genes? How do you know a priori genes are not informative? Or arbitrary if the sum of the gene counts across all the samples is less than 5, as we did, we discard it?**

    Ash: We are only going to remove genes with very few reads, so yes an "arbitrary" low number e.g. 5 counts (you would maybe adjust this depending on your read depth). These counts are not reliable as they are really in the region of random noise. With other packages such as EdgeR you have to be more stringent with your filtering and identify uninformative genes, but we will see tomorrow how DESeq2 will do this for us later on. The rational with other methods (e.g. EdgeR or limma) is that we will have to apply a multiple testing correction to our statistical analysis and so the more genes we test the less powerful our analysis and the less truly differentially expressed genes we will

be to detect, however, DESeq2 will do this for us. In this workflow the removal of the low count genes is more to clear up the visualisations.

70. <Fernando><How I can save my script on my computer? I can save it only in the remote desk>
    Paul: To download any of your files via the Rstudio server follow these instructions:
    - Select the "File" tab from the set of tabbed windows in the bottom right
    - Find the files you are interested in and click the checkbox next to them
    - Once all the files you want to download are selected click the "Export…" button at the top of the "File" tab. This button may be hidden in the "More v" dropdown menu
    - R should then prompt you to name the file to be downloaded

    <Gabriel> I cannot see where the "More v" dropdown menu is (?)

    <Paul>

71. <Fernando> Yesterday, we only worked with MCL1 DL, if we want to analyze all of them, should we do one each time, as we did yesterday?>
    <Abbi> If you remember yesterday we did an example where we counted lots of bams in the same command and it made a large output table with a column for each bam. For that example we just did a set of small bams so it would run quickly for you but today Stephane has loaded in the count data of all the fullsize bams run together.

72. <Shaimaa> what does rlog mean?
    <Sankari> Regularised log transformation. It takes into account the size factor normalization (library size, etc) and makes log transformation of your data. It outperforms the other methods when you have a small number of samples (n< 30).

73. <Shaimaa> can I use the rlog transformation for proteomics data as well? And overall can I apply all these concepts to proteomics data?>
    It depends on what type of proteomics data you are referring to.
    <K> Yes but in proteomics it depends if it's count data or intensity data, you need to modify the concept accordingly.

74. <Komal><What is the PCA analysis based on? I understand it separates the samples according to the expression pattern but what is the input rlog is taking to plot?>

Ash: The rlog takes our raw read counts and normalises for library size and takes the log2. This puts our data into a distribution that is closer to normal before we carry out the clustering.

75. **<shaimaa><Why we convert the datacount to matrix?>**
    Ash: This is just an R object thing. Our original count table object is a `data.frame`, but the input for the prcomp() function requires a `matrix`, which forces all the columns to be the same type of data (in this case numeric). In a data.frame we could have some columns with numbers and others with character data, e.g. gene names. This is a way for prcomp to ensure you give it the right type of data.

76. **<shaimaa><Just to make sure that I understand right could I interpret the PCA as following :-**
    -the main difference due to the cell type and account for 62% of the variation and difference between status account for only 14.73% of the variation
    - Effect of the status is not the same between different cell types as distance between status is not the same
    Ash: Spot on!

77. **Q: What is the <- read_tsv("data/GSE60450_Lactation.featureCounts", comment="#") what comment means#??**
    Ash: It tells the read_tsv command that there are comment lines in the file that start with a '#' and it should not read these lines. The featureCounts output that we generated yesterday starts with a line that describes the command that we used to generate it. This is prefixed with the '#', the table of counts starts on the next line.

78. **<Esther> I have always had the next question: How do you know "the real hypothesis"How do you really know what is the behaviour of your data? . You do not know it, that is why you are doing the analysis>**
    Ash: Sorry, Esther, I am not clear what you mean by "the real hypothesis". We have a question here: is the gene expression different between two sample groups. We can define a Null Hypothesis (H0): there is no difference in mean gene expression and an Alternative Hypothesis (H1): there is a difference in mean gene expression. We test the Null Hypothesis and reject it if the test statistic is extreme and instead accept the alternative - there is a difference in means.

79. **Esther Where are you getting the data or the results to calculate the size from?**
    Ash: Do you mean the data for doing a sample size/power calculation, in general terms?
    Esther yes, sorry

Ash: So, you need to have some prior data, ideally this would be a pilot study where you can test the detection limits on the gene expression under the different conditions of your experiment for genes in which you might be interested (could be all of them if you are on an exploratory study). Alternatively, you might find public data with similar conditions. In practical terms, due to the cost of RNAseq experiments, both in terms of money and time, it is simply not possible to run a pilot study, and often there is no similar experiment (with good data) in the public repositories, and so we have to base our sample size by generalizing from previous experience/data. Furthermore, pragmatically, we rarely have anywhere near the number of replicates that would give us the power we would like to have, so the general rule of thumb is as many as your budget/resources/time can tolerate. The absolute minimum always being 3 otherwise we cannot model the distribution. e.g. for mouse experiments such as this we tend to recommend 6 replicates per sample group, however, if we get 4 we consider that a good situation. Unfortunately the statistically ideal experiment has to compete with our real situation.

80. **<Gita> <Why is the Gene Expression Data Matrix transposed for most analyses? We always do this before PCA, also for Co-Expression, as well as DESeq regression. We put Genes usually as rows and samples as Col, but then we transpose this for calculations. >**
    Ash: Basically, this is just a function of the tools you are using and the way the author of those tools decided the input should be. In general older tools use the wide format like prcomp which is to have the genes in columns and the samples in rows, and more recent tools favour the long (tidier data ;)) format with the samples in columns and the genes in rows.

81. **<Esther><Probably it is a stupid question but: If we can do with R the same than we did yesterday, Why we did it?...ok, thanks >**
    Ash: Yesterday we were working with the full sequence data, the files are very large and many of the analyses we were doing were very computationally intensive. R does not handle high memory tasks well and we really need to use a more powerful means of doing this. There are ways to do some of things we did yesterday with R, e.g. the featureCounts, but it is much much slower.
    <Justina> To follow up on this, what shall we use for our own data analysis? Terminal on Macbook?
    <Abbi> This was slightly answered in question 39 but to summarise, you can install the programs from yesterday on your machine to use but you may want to check what computational infrastructure is available at your institute, as your macbook may not be powerful enough or have enough space to handle this level of processing. The R parts should work on most personal machines but the processing from yesterday will most likely take more.

82. **<Romit> <If we are seeing batch effects in our PCA plots ie each set of expression data are distinct based on when they were conducted, not just the**

experimental conditions, can we add a term for batch in our design matrix to account for this? >

Ash: Yes, this is one very good way of accounting for batch effects. The situation you describe is something we often encounter. If you have just two experimental conditions, e.g Control & Treatment, the linear model equates to to a t-test, then adding the Replicate as an additional parameter essentially turns it into a paired t-test. It is still possible to add the Replicate if necessary in more complex situations. We recommend detailed and careful recording of everything during sample generation so that the source of batch effects can be identified and if possible added to the model. That said this is not always possible, particularly if your batch effect is confounded with an experimental factor. The best thing is to alway consider potential sources of batch effects during the experimental design and try to avoid or control them.

83. <Fernando <Is negative binomial regression , something similar to 2 to the power of DeltaDelta Ct in q pCR?>

    Sankari: Formula might look similar but it depends on the distribution of your data results after qPCR. It can get to Poisson or NB.

84. <Charlotte> <What sort of experiment would we use a model matrix without an intercept for? Is it when we don't want to compare to a reference sample e.g. treatment 1 and treatment 2 to control, but rather we may want to compare treatment 1, 2, 3 to each other?>

    Ash: Actually, it doesn't matter, all experiments can be modelled with design either using an intercept or not, it is simply how easy it is afterwards to extract the contrasts that we want. In terms of analysis using DESeq2 we always use the intercept design. In your second example we would maybe set treatment 1 as the intercept.

85. <Holly> <Will we cover how to actually perform these statistical tests in R ourselves?>

    Ash: Yes, we will do this in practice tomorrow morning.

86. <shaimaa> <how we get the stat column and what does it indicate>

    Ash: we will cover how to do this in R tomorrow. It is the test statistic which is what we use to generate the p-value. All of this is done by DESeq2, but it is important to have an appreciation of the underlying principles.

87. <shaimaa> <sometimes i got the p value distribution with a very small peak around 1 is it okey or what shall i do>

Ash: We will cover this tomorrow. There are many possible reasons and we will show you a website that talks about some of them. Usually this means that you have a bimodal distribution. In RNAseq this can be caused by a group of genes that are strongly expressed in one condition but totally absent in all others, however, this is rare. More likely is that there is some problem in the upstream data preparation.

88. **<Tom Dennison> <What is the purpose of using Log2FC rather than just FC?>**
Ash: You can use either the log2FC or the FC in your interpretations of the results, however, for visualisation an upregulation of logFC of 1 and downregulation of logFC -1 are easier to represent then the equivalents of 2 and 0.5:

In log2FC: -2……-1……0…...1…...2
In FC : 0.25..0.5……1……..2…….....(3)……..……..4

We'll see tomorrow with some of the plots. (it is indeed a transformation aiming to optimise graphical representation [Dom])

89. **<Fernando> <About question 71, Could you give me a clue, what was the step where we filtered or chose that subgroup?>**
<Stephane> wrt 71, MCL1 DL is a sample. Is that what you mean by subgroup?
<Fernand> yes

90. **Q: I'm not sure I understand what do you mean by pilot study to decide the number of replicate ?>**
Sankari: A pilot study is a study in small size which can replicate your exact main experiment but with a small number of replicates. The results from this study will be used to model the variations you would get from your replicates. Using this, we can calculate the sample size needed to achieve high statistical power. If you are performing a completely new study where it's hard to know how many replicates you need or you are practising a new technique, a pilot study would be useful. Most of the in vivo experts prefer having a pilot study so that the main study goes without any issues.

91. **<shaimaa> I just want to ask what is the best way to exclude some gene (from gene list) from GTF file and make the output the same format as GTF so gffread can easily read it as when I use fgrep -f I end up having gtf file as output but I wasn't able to convert it to fasta or gff3 by gffread so I end up load it to the R by read.gtf function and convert it to fasta on R by compare it with fasta file of original GTF Im not sure if this is right or what the wrong thing I have done‹**

92. **Q: Regarding question 73 about if we can treat the proteomics data as rna seq data you said it depends if the result is intensity or count I think I have intensity which indicates the count right? How I convert intensity to count and actually how I make sure that it is count or intensity**
Chandu: The purpose of rlog or other transformations like vst, is to remove the dependence of the variance on the mean. They equally work well for intensities or counts data. Once you transform the data visualise mean vs sd values. Intensities are continuous values (like, 1.0, 1.1, 1.2, 1.3 etc), but counts are discrete and have values in integer form (1, 2, 3, etc).

Day 3 Questions

93. **Q:Fernando: We always need to create this file (<- read_tsv("data/SampleInfo.txt")) to add information for the file with counts?**
Chandu :It is ideal if you create sample metadata (SampleInfo.tx) and load it for DE analysis. You can also create on fly, but this process can be error prone. Therefore create metadata file and counts file separately, then use them for analysis.
Q Should I create with txt editor?
Chandu: Yes you can, but again, manual creation is error prone. You have to be extremely careful.
Ash: To be clear, it is fine to create your "sample sheet" with the sample information (meta data) in text editor or spreadsheet. But be careful that you enter data correctly as mistakes will propagate through the analysis. Also, once you have your original version, save new versions for any changes that you make.

Q. I am bit lost, so I do not remember in which step we generated or when we can generate this file?
We didn't generate this file, we just provided it. It contains all the meta data from the experiment, essentially it is a description of each sample.

94. **<Fernando> <Why I had this error, how can I remove NAs, I used "preprocessing.RData">**

```
> ddsObj.raw <- DESeqDataSetFromMatrix(countData = countdata,
+                                      colData = sampleinfo,
+                                      design = design)
converting counts to integer mode
Error in DESeqDataSet(se, design = design, ignoreRank) :
  variables in design formula cannot contain NA: Status
```

Chandu:One quick solution is using na.omit function, something like below. countdata <- na.omit(countdata) # remove all the lines with NA values.

Zeynep: I saw this error with another student, and the problem was due to a typo at the factoring step (sampleinfo$Status <- factor(sampleinfo$Status, levels = c("virgin", "pregnant", "lactate"))) - can you re-load Robjects/preprocessing.RData and check your command for typos.

95. **<Romit> <Why does the non-normalized data from an RNAseq experiment not centre around zero by default? What's the cause of this natural skew / quirk? >**

Ash: This is mainly due to library size difference. If I have a gene that has the same expression level in two samples, but one of them is sequences to twice the depth, then it will appear in the raw data that it has twice the expression in one sample.

96. **<Justina> <A question about normalised counts. I am comparing immune cells (the same cell but of different origin) at different activation status (mock vs activated). Of course, activation drives the cell to dramatic extremes. Should I be worried about normalisation?>**

Chandu: What percent of genes you think are significantly differentially expressed?

<Justina> Not sure the exact percentage but about 500 genes, so I guess no - it is not as dramatic as I thought.

97. **<Komal> <Why are some padj values NA?>**

Zeynep: This is due to independent filtering in DESeq2.

Chandu: To add on, if all the samples in a row have zero counts then the baseMean will be zero, then you get NAs. Another possibility is that if within a row few samples are extreme outliers, then you get NA.

98. <Alyce> <When making the additive model I got this warning, does it matter?

```
converting counts to integer mode
Warning message:
In DESeqDataSet(se, design = design, ignoreRank) :
  some variables in design formula are characters, converting to factors
>
```

>Abbi> No that's fine, it's just letting you know. You can raise your hand and get one of us into a breakout room if you need more help.

99. Q plotMA is showing plot for only one sample, can we get that for all samples?
<Stephane> You would need to make one plot for each sample, combined in a multi-panel plot. That was one of the challenge in the version of the course given in a class room (the session on preprocessing and preparing the DESeq object). We took it out for the online version. The 'in class room' version of the course is available to students on the github site, see the 'extended' version. It may not be just now because the course is not finished.
<Abbi> The previous version is available now in the extended materials section at the bottom of the website.

100. <Tom D> <For a model with two factors (e.g. `~ CellType + Status`), when comparing e.g. pregnant vs lactate is this comparing *all* pregnant vs *all* lactate ie the intercept is samples from all 3 cell types? So the intercept is different for different comparisons? Or is the intercept just virgin/basal samples?>
<Stephane> The intercept remains the same for a given model, eg `~ CellType + Status`, regardless of the contrast (coefficient) of interest, eg pregnant vs lactate. The intercept represents the mean for the samples in the first level (baseline) for each of the factors in the model. So here if CellType is a factor with levels basal and luminal, in that order, and Status a factor with levels virgin, etc, in that order, then the intercept will indeed be for the basal cell in virgin mice. One can set the order of levels for a given factor when assigning the variable, with factor(, levels=), or if the variable already exists, with relevel().

101. <Fernando> The link Github repository (source material) in the google doc is broken
Chandu: https://github.com/bioinformatics-core-shared-training/RNAseq_May_2020_remote
Ash: Thanks Fernando, it is, I will fix it later. There is another link right at the top of the page in the banner that is working.

102. <Romit> <Can't we just redefine the reference factor in R before fitting the model to extract the contrast we want? E.g.

```
sampleinfo$Status <- factor(sampleinfo$Status,
                            levels = c("virgin", "pregnant", "lactate"))
```

sampleinfo$Status <- relevel(sampleinfo$Status, ref = "pregnant")
Which then sets the intercept term to pregnant >
<Stephane> Indeed, you can, see question 100 above.

103. <Romit> < In other branches of stats you might fit a complex and simple model and then compare them with an F-test which would output a statistic to tell you if there was a significant difference between them. Does LRT not have some Bayesian confidence output that tells you the same thing but incorporates information penalties?>
Ash: I'll leave Dom to answer the stats question, but I just want to point out that here we are fitting the models 25000+ times, once for each gene and each and every gene has different fit, and when comparing the two models, for some genes the simpler model will be adequate and for other it won't. It's difficult to come up with an overall test that would then tell you which to use for the whole gene set, some heuristic approach based on the biology and the question at hand is required to make that decision.
Dom: Likelihood ratio tests (invented by Fisher, btw) are typically purely based on the ratio of the (maximum) likelihood of a restricted and a full model (note that REML [restricted maximum likelihood] led to biased LRT in some models) and is therefore more "frequentist". Bayesian add prior to the equation and would then consider the posterior odd ratio of two models. Nothing incorrect there assuming the priors are well defined but a computer cost: MCMC would likely be required to compute the test of interest, which could prove challenging given the 25K tests mentioned in his answer

104. <Shaimaa> <Shouldnt we also do the DEseq for the new design interactive model to compare it with the previous one as we didnt do it>
Ash: No, you can proceed straight to the LRT test, the modelling is still being done, the difference is just the test being applied to generate the p-values. In this case the LRT test is being used to compare the 2 models, for DGE we used the Wald test.

105. Q: here in exercise 2 how we call the more complicated model as the reduced?>
Ash: It is reduced with reference to the model we are testing it against. Here were are testing ~CellType + Status + CellType:Status (interaction model) with a simpler (reduced) model ~CellType + Status.

106. <Gabriel> Can we just focus on the genes where the interaction model fits well, and use the interaction model to analyse only those genes?

107. Q: Can we simply apply DEseq function to the Proteomics intensity data?>

Ash: Actually no. The reason that the RNAseq analysis uses a negative binomial model is that the data are count (i.e. integer) data. In contrast microarray data, back in the day, and proteomics intensity data come on a continuous scale. There are both (approximately) normally distributed (once log transformed) and so we need to use a different model. In this case proteomics intensity data looks very much like microarray data and many of the same analysis techniques are appropriate. You can have a look at the Bioconductor package qPLEXanalyzer to see how limma can be used with proteomics data (full disclosure: I am one of the maintainers of the package, other packages are available).

108. **Q: I tried to apply colum_to_rownames but didnt work**

```
> raw.pto<- raw.ptoo %>% column_to_rownames("Gene.ID")
Error: `.data` must be a data frame without row names.
Run `rlang::last_error()` to see where the error occurr
ed.
>
```

>

Ash: I am unable to say for sure what the problem is as I can't see where the raw.ptoo object came from or what it is. The error message is telling you that it needs to be a data.frame in order to run 'column_to_rownames' on it. I am guessing that it is a DESeqResults object. If so, you need to change it to a data.frame first using 'as.data.frame'.
Stephane column_to_rownames, with a 'n' in column?

109. **Q: <If im dealing with non model organism thats not in biomart or ensemble however I have csv file with annotation How to annotate>**

Ash: There are various ways to merge two tables. You need to read you csv file into R using "read_csv". Then, assuming you now have an annotation table called "annot" with a column called "GeneID". You could do:

```
library(tidyverse)
  resLvV %>%
     as.data.frame() %>%   #see answer to above question
     rownames_to_column("GeneID") %>% # name the column the same
     left_join(annot, by = "GeneID")
```

We will actually do this with the annotation table we are generating in the course.

110. **<Fernando> <i am still struggling with question 93>**

A: Fernando, let's chat in a breakout room, please send me a message in the chat.

111. Q:<How to Annotate to other organism orthologue if this make sense? If I want to see the human orthologue of my genes for eg>

Ash: There are a number of ways you could do this. One of the easiest would be to pull this information from ensembl via Biomart. You could do this through their web interface, but there is also an R package biomaRt which allows you to access the databases directly. We normally use this for annotation and it is explained our Extended materials. You would need to find the orthologues column names, and this also explained.

112. <Jacob Dundee> Q: I tried running the most recent code to produce annotLvV however it came back with the error
Error in rename(., logFC = log2FoldChange, FDR = padj) : object 'log2FoldChange' not found
Would anyone be able to assist in why this error occured?
Stephane ''log2FoldChange'' may not be the correct name for the column storing the information. Use colnames() to find the actual names of the columns, or click on the object name in the environmeny tab to view the object

113. Q: <If I want to remove certain genes from GTF file before doing analysis this is mainly when Im doing noncoding RNA prediction so I wannt to remove the genes have ORF or protein coding potential etc how can I do it
N.B I tries to do it in linux by cat annd fgrep -f but the output is somehow is weird as for example I cannot convert it to fasta or gff3 by gffread could you help?>

Ash: Yes, you would want to do this at the command line using e.g. grep. I have done this many times, it should work fine, it's going to be down to the command you used.

114. <Tom D> <With a model with two factors (e.g. `~ CellType + Status`) would a comparison (such as lactate vs virgin) give the same results as if you extracted only basal cell/lactate and basal cell/virgin samples (i.e. 2+2 samples)?>

Ash: No, this would only be using half the data and would give different results.

115. <Romit> <Is there a standard way of managing multiple gene annotations between Ensembl IDs and ENTREZ?>

Ash: The short answer is 'No'. There's no standard way to do this. You will have to make decisions based on what you want and what is practical. You could manually curate them and choose the ENTREZ id that makes more sense to you, you could randomly select one or the other, you might keep them all (e.g. concatenate the ENTREZ ids with a ';'), or you might decide to throw them out. You'll have to decide how important things are to you. If you have only 10

genes with multipel ENTREZ maybe manual curation is okay, but if it is 1000 then maybe not. Unfortunately there is not really a more helpful answer.

<Romit>*<as an extension to this, do we need ENTREZ id's at all?, can't we just stick with the ENSEMBL and gene symbols. The ENSEMBL ID does not tell us which exon we are looking at so we can't relate them back to the 'correct' ENTREZ ids***?>**

116. **<Name> <regardinng 113 what could make the GTF couldnt be read by gffread I tried it several times it never worked ?**

Ash: Sorry I can't really answer this without a detailed look at the GTF and some playing with gffread (never used it). Maybe of of the other trainers might be able to help.

117. **<Georgia> <Why do we carry out the p value histogram check on the unadjusted p values as opposed to the FDR values?>**

Ash: First the FDR will include NA, second this is a pattern that is a consequence of the p-value distribution - for all H0 the pvalues are uniformly distributed 0-1, but the H1 pvalues tend to 0, so we get this pattern. The FDR doesn't work as well as it will depend on the multiple testing correction used and doesn't necessarily have this distribution.

118. **<Yunjia> <How to do ddsShrink if my coef is not in the resultsNames(ddsObj)? For example the contrast parameter we provided by ourselves>**

Ash: You can extract the results table first and then run the lfcShrink command on that.

119. **<Charlotte> <What is the purpose of lfcShrink? I missed the explanation sorry.>**

Ash: When doing downstream analysis we are concerned with both fold change and FDR/p-value. The fold changes for lowly expressed genes - genes with low counts - tend to be inflated due to noise. This tends to bias our impressions from plots and also the results of functional analyses based on ranks of fold change towards these genes. The shrink uses the information from the entire dataset to shrink these lower expressed genes' fold changes to reduce their overall influence. I recommend reading the DESeq2 manual's explanation of this.

120. **<Tom D> Re Q114, so in this situation, "lactate vs virgin" compares *all* lactate (basal + lumenal) vs *all* virgin (basal + lumenal)? Even though the intercept is virgin/basal (see Q100)?**

Ash: Yes, this is correct for the additive model. To extract the difference between basal and luminal for virgin, the difference basal and luminal for lactate etc. then you would use the interaction model. If there is no interaction these would be the same if the experiment was sufficiently powered, however, with just 2 reps sampling error would make this unlikely.

121. **Q: If I understand well the shrink function will remove the low logFC genes but how it decides the cuttoff value?**
Ash: No, it is not a cutoff. It is shrinking the logFC of genes with low counts. Please see question 119 above.

122. **Q: MA plot → I dont know what is the message it gives?**
Ash: The X axis is the average expression of the each gene across the sample, the y is the fold change between the two groups. Gene that are far from the horizontal logFC=0 are more differentially regulated, you can get a sense of their importance from their overall expression. A strongly differentially regulated gene with low expression may not be as important and a highly expressed gene with a low differentia regulation. The plot allows us to asses the overall affect of our treatment/contrast across the whole transcriptome. E.g if you had an extreme effect due to a treatment with some drug you might see and overall suppression of gene expression - the cloud would be more skewed to the bottom of the plot.

123. **Q: Do I have to do lfcShrink???? I really think it could mask some results?**
Ash: For visualisation and for ranking genes for functional analyses such as GSEA it is highly recommended. Random noise that has a lot of effect at the lower end of the gene expression where there are few counts for each gene causing inflation of the fold changes in this region. If you are particularly interested in lowly expressed genes then perhaps you would not want to, but even then you have to be very careful about over interpreting the fold changes. If you have particular interest in very lowly expressed genes it is recommended to increase the sequencing depth in order to increase the signal to noise ratio. I strongly recommend that you read the section in the DESeq2 manual where they discuss shrinkage.

124. **<Tom D> Sorry one more re 114 and 120. So would an interaction model for a specific comparison (eg basal cell lactate vs pregnant etc) give the same results as extracting just those samples? Or does the model use the other samples to increase power somehow?**
Ash: Yes, DESeq2 will "borrow" information from the other samples, both in the estimateSizes and the estimateDispersions steps and the modelling, so the results will be different. In an ideal world the fold changes would be the same, the p-values would change, but the influence of sampling variation within sample groups means that they will end up slightly different.

125. **Q: how comes the heatmap includes data from all conditions (luminal, basal, virgin, lactate and pregnant) when we started with the results table for lactate vs virgin?**

<Stephane> because the results table for lactate vs virgin informs on the genes to display expression for, while the object keeps expression levels for all samples, not just those included in the lactate vs virgin comparison

126. **Q: If you have two conditions e.g. cell type and status, and interaction is present but you do not have enough replicates to use an interaction model, is it valid to subset the data (e.g. just have luminal pregnant and luminal lactate present in the count data) and then run pairwise comparisons on the data subset with DESeq2 (in this case luminal pregnant vs luminal lactate) or should you run an additive model with the full dataset and then extract the comparisons that you are interested in?**

Ash: If you don't have enough replicates to power the interaction model, you wont have enough to power the simple analysis either. E.g here we would only have 2 reps per group, which is not ever sufficient in a simple contrast. As a general rule an interaction model will be borrowing between our sample groups and this adds power. E.g if you do a large time series with two conditions, it might be that 2 samples per a group will be sufficient as there are lot of samples. I would generally not recommend this as a plan for an experiment as from a practical point of view you risk the whole experiment if a few of your samples are lost due to technical mistakes in processing.

127. **Jacob D: How long will the course page be available with all the resources after this session? Thanks!**

Ash: Should be available 'forever' where 'forever' means as long as GitHub exists. I think this should be well past your need to refer back to it ;)

128. **Anna: Is it possible to apply GSEA to single cell data between clusters?>**

Ash: I should think so, but I'll get Stephane to come back to this as he is more experienced with scRNAseq analysis.

Abbi: Yes, in a very similar way to for bulk analysis. I suggest reading chapter 12 of https://osca.bioconductor.org for more information.

129. **Q: Leading from Q126, just to check I understand, if you have two replicates per condition but many conditions, is it best to use an additive model with the whole dataset? As I assume an interaction model would not be advised based on the low replicates.**

Ash: This is hard to say for all situations. It depends on your experiment and the response to the different conditions. If you are unsure I would suggest talking over your data with a statistician or at least a lowly bioinformatician.

130. **Q: Sorry, last addition to Q126, is it generally not advisable to subset data before running DESeq2, especially with low replicate amounts? Instead always incorporate all data within a model**
A:This is a bone of contention between biologists/bioiformatics analysts and statisticians. Statisticians would generally say always put all the data in and build the appropriate models. I would say only split the data if the sample groups can be considered as separate experiments. I know other methods that will subset data in different ways and do different DESeq analyses for each contrast required - I think this is very wrong. But this is biology and not physics and sometimes we need to be pragmatic.
Ultimately, if you are looking at general trends (e.g. pathway enrichment) and the method you use strongly changes your biological inference at the end of the day, then you might want to question the experiment itself. If you are looking for specific genes then you need to you need to have a very stringent experimental design in order to be confident the gene you've got are reliable - if your gene expression response is strong this is easy, if not then you need lots of reps.

131. **<Fernando> Single Cell RNA Seq, follows the same data treatment?**
Ash: No scRNAseq data requires different methods. You should look at the workflows either for scran/scater or Seurat, which are packages specifically for scRNAseq.
Abbi: I would start with this manual https://osca.bioconductor.org its the clearest I have found for teaching yourself how to do scRNASeq analysis. The Seurat website does also have lots of workflows but I don't think they are as beginner friendly.

132. **Q: Is the green line indicative then of the directionality of the fold change? I.e where it crosses 0 is when it goe from positive fold change to negative?**
Ash: No the green line is the cumulative enrichment score. The way the enrichment score is calculated is this: imagine walking along the the black line at 0 from left to right. Each time there is a tick mark, which indicates that that gene is present in the pathway, we add to the enrichment score a bit (the green line goes up), each time there is no tick, i.e. the gene is no in the pathway, the enrichment score is decreased and so the green line goes down. The whole green line shows you how the enrichment score varies from left to right. The final ES is the maximum (absolute) that the green line hit. This does take a bit to get your head round, the GSEA website has plenty of help pages and explain this all in more details.

133. **<Alyce> My table has overlapped the image as Stephane's originally did**
Ash; Sorry this appears to be just a bug in the RStudio server. Try clearing all of the plots using the brush icon and rerunnning the table plot. That worked! thanks

134. **<Gabriel> I got this error, but I copied from the on line material:**

```
> topPathways <- fgseaRes %>%
+   top_n(20, wt=-padj) %>%
+   arrange(-NES) %>%
+   pull(pathway)
> plotGseaTable(pathwaysH[topPathways],
+               rankData,
+               fgseaRes,
+               gseaParam = 0.5)
Error in lapply(pathways, function(p) { : object 'pathwaysH' not found
>  GSEA - gen
```

Ash: It hasn't found the pathwaysH object, have you loaded it? Please ask for help via the chat.
load("Robjects/mouse_H_v5.RData")
pathwaysH <- Mm.H

<Gabriel> Yes, I did load it:
 # load pathways
load("Robjects/mouse_H_v5.RData")
pathwayH <- Mm.H

<Gabriel> Found a typo, I wrote pathwayH instead of pathwaysH…
Ash: No problem, Gabriel. 90% of errors when writing code are traced back to typos. The <TAB> autocomplete is a lifesaver in this respect.

135. **Q: Please I dont understand how we knew if these pathways are actually overexpressed or its downregulated ???**
When I do the GSEA I took the upregulated genes and put them and got the list of pathways and then I redo it with the downregulated genes. Is this wrong?
Ash: There are various ways that are all perfectly legitimate for ranking your genes and running GSEA. I have never actually come across someone splitting the data into up- and downregulated genes first though. If we rank the genes using the FC in the ranking statistic then we know all of our upregulated genes are at one end and all the downregulated gene are at the other end. This way we can tell about the preponderance of direction in our pathway. However, this is not always valid, for some experiments in some pathways we might expect the effect on a particular pathway to downregulate some genes and upregulate others (actually this is most commonly the case, but often the overall balance is one way or the other) in which case you might want to ignore direction and just use absolute fold change or just the p-value.

136. **<Tom D > Does the green line always end up back at 0 and if so why?**
Ash: Yes, and this is one of the harder questions I always hope doesn't come up. From my explanation above you would think should not be the case. The algorithm adjust the up and down increments of the ES as it goes. I can't really

explain it succinctly so I would suggest the best thing is to read the original paper or the documentation on the GSEA website.

137. Q: NES when its negative what does it means ?
Ash: It means that most of the genes hitting the pathway are clustered towards the right of the plot. Ie. if we order from highly expressed to lowly expressed, then the general trend in the pathway is for downregulation of genes

I

138. Shaimaa What sign in ranks means?
Ash: it means we take the "sign', i.e. -ve or +ve, of the log2FoldChange. This results in us ranking by most significantly up regulate to non-significant to most-significantly up-regulated<Tutor>

139. Charlotte W> is there a way to combine conditions when testing contrasts with DESeq2 e.g. you want to know which genes are differentially expressed in brain neurons vs two other neuron types, to look for genes specific brain neurons as opposed to neurons in general
Ash: The easiest way to do this is in the sample sheet by creating a combinatory category and then using that in the model. E.g:

SampleName	CellType	CellGroup
Sample1	type1	TypeA
Sample2	type2	TypeB
Sample3	type3	TypeB
Sample4	type4	TypeB

Then use "~CellGroup`
You can do nested analyses I believe, but I have never had to so not sure how you would do it. I think I have seen discussion about it on the Bioconductor website.
Charlotte W: Thanks so much!

140. Q: I used the http://www.webgestalt.org/ for overrepresentation analysis and here there is no rank column to be considered so thats why I took the upregulated list and downregulated separately and see which pathway are over annd which are under is this wrong
Ash: I am not familiar with webgestalt. There are many different overrepresentation tools available and without understanding the algorithm it is hard to say. I think the important thing here is to remember that these methods are indicators of what might be interesting rather than answers to questions. They give and overview, so many different approaches can be valid.

141. Q: Im dealing with non model organism so I took the human orthologue and do the enrichment as if its human?
This is one way to do it. In fact the mouse pathways we used for the GSEA were generated this way, the Broad only curates human lists. I used to work with a lot of weird non-model organisms and this is really the best you can do, you just have to view any results with the appropriate amount of skepticism. The other option, if you can, is to generate your own gene list of interest from literature or your own research (obviously not from the same experiment!). e.g. if you were investigating development in your organism you might collate lists of genes known to be key at different stages of development and then test enrichment against them.

142. <Robyn> Is there any possibility of having access to the remote interface tomorrow too? There has been a lot of information today and would be good to have some time to digest it all!
Paul: I'll give a little thought to whether this is possible and make an announcement in a few minutes during the wrap up
Robyn: Thank you!

143. <Charlotte><In the last pathview example, why do we use the non-shrunk logFC for our diagram?>
<Stephane> shrunken LF should be used. Given that genes with high LFC (either direction) and low read counts are unlikely to have been selected as differentially expressed, they would not have been included in the KEGG over-representation analysis. For the genes included, the difference in LFC before and after shrinkage would be small.

144. <Dawei Sun> Not entirely sure how the pvalue in GSEA analysis was calculated?
Ash: GSEA uses a permutation approach (Dom our statistician loves these). Basically, do the analysis for our data. Then randomly reorder the genes and rerun to get a new ES. Do this 10000 times and you can then calculate how likely it is to get the ES for our ranking if the ranking is really just meaningless (random).

145. <Gabriel> I am getting this error: (for some reason it cannot find pathview function)

```
> browseKEGG(kk, 'mmu03320')
> browseKEGG(kk, 'mmu04060')
> sigGenes <- shrinkLvV$FDR < 0.01 & !is.na(shrinkLvV$FDR)
> logFC <- annotLvV$logFC[sigGenes]
> names(logFC) <- annotLvV$Entrez[sigGenes]
> pathview(gene.data = logFC,
+          pathway.id = "mmu040060",
+          species = "mmu",
+          limit = list(gene=5, cpd=1))
Error in pathview(gene.data = logFC, pathway.id = "mmu040060", species = "mmu",
  could not find function "pathview"
>  GSEA - gen
```

Ash: Try `library(pathview)` first. It think this is missing in the materials. I'll check and if it is we'll amend

Gabriel:
Now it seems I have loaded the 'library(pathview); however, cannot visualise the pathway with the genes up and down regulated. What's the script to visualise the KEGG pathway?
Stephane:

```
library(clusterProfiler)
library(pathview)
sigGenes <- shrinkLvV$FDR < 0.01 & !is.na(shrinkLvV$FDR)


logFC <- shrinkLvV$logFC[sigGenes]

names(logFC) <- annotLvV$Entrez[sigGenes]

# write plot to file in working directory,
# check with getwd()

pathview(gene.data = logFC,
        pathway.id = "mmu04060",
        species = "mmu",
        limit = list(gene=5, cpd=1))
```

146. **<Alyce> I got the same error as above ^**
<Stephane> I had forgotten to load the pathview package. Use library(pathview) before calling pathview()

147. **Sudipta : There seems to be a problem with the github link, it says page not found.**

<Abbi> The link at the top of the page should still be working.

148. <Regarding Q141 thats really amazing suggestion could you help me how to do enrichment against specific list??????
Ash: If you wish to use fgsea as we did you just need to create a list object like the Mm.H object where each item in the list is a set of genes of interest:
myListofPathways <- list(myPathway1=c("gene1", "gene2", "gene3"),
myPathway2=c("gene1", "gene4", "gene7"))

The gene names need to match those in your data.