Next Generation Sequencing Technologies: Data Analysis and Applications

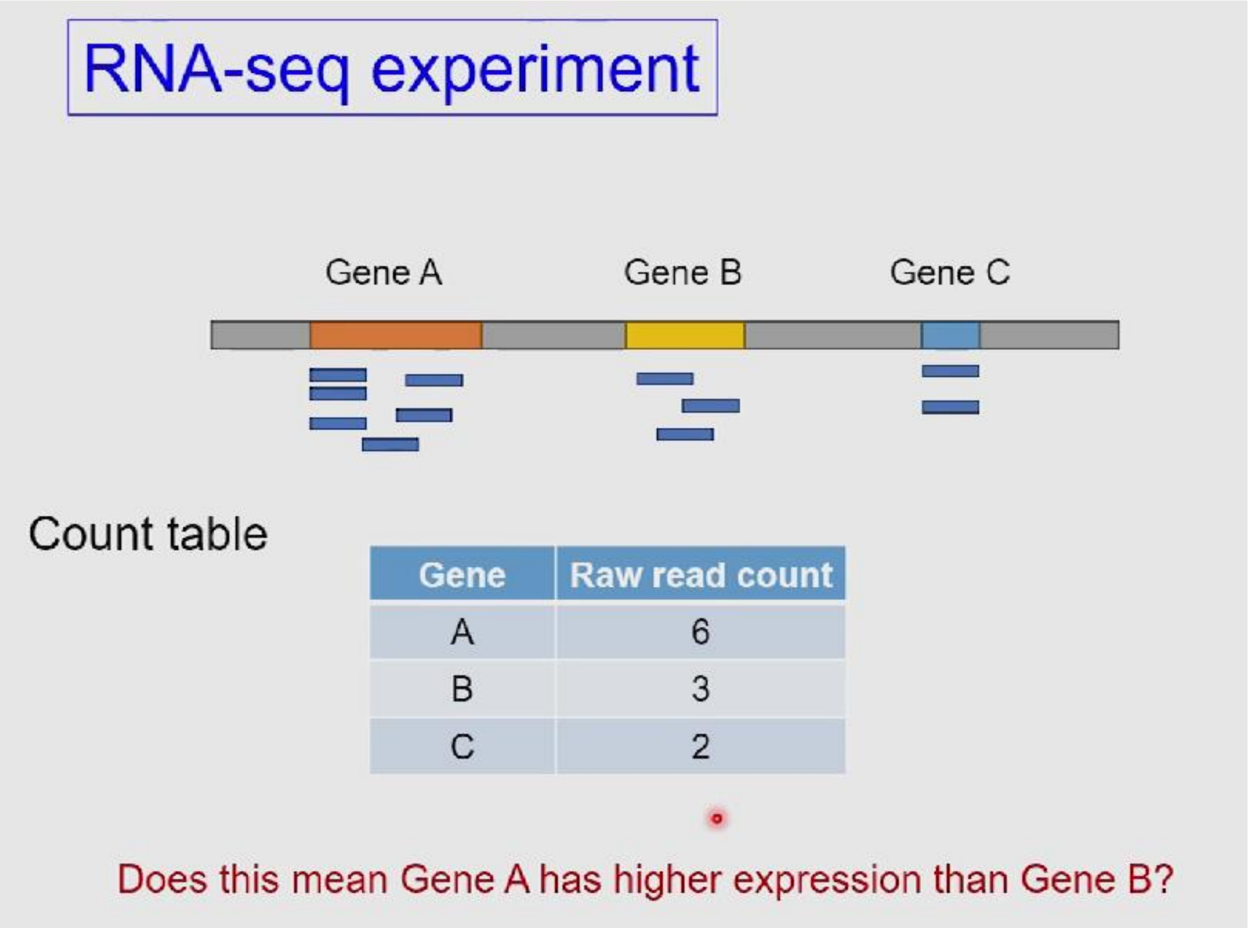
Biases in RNA-seq experiments

Dr. Riddhiman Dhar, **Department of Biotechnology**

Indian Institute of Technology, Kharagpur

Good day, everyone. Welcome to the course on Next Generation Sequencing Technologies, Data Analysis, and Applications. In the last few classes, we have talked about the transcriptome assembly process, and then we have talked about the transcriptome quantification process. And in between, we had a very short hand on how to get to the transcript abundance part or the raw count part from the read data. So, we will talk about the biases in RNA sequencing experiments in this class. So, we touched upon this topic very briefly in the last class when we were talking about the alignment-free methods and we will talk about these biases in a lot more detail in this class.

Now, this is very important because these biases can affect our quality of analysis and our quality of inferences that we draw from these RNA sequencing experiments. So, we need to understand these different types of biases that are present there. Sometimes we may not get to the exact origin of these biases or why they arise. So, this is something we still cannot predict, but we observe these biases in the data, and this is something that we need to correct before we can proceed with further downstream analysis. So, these are the concepts that we will be discussing in today's class. So, the biases in RNA sequencing experiments and methods for bias correction are okay. So, let us look into these different types of biases, and here are the K words. So, we will talk about positional bias, sequence-specific bias, and finally, we will introduce the term normalization. So, let us start with a very simple example of this RNA sequencing experiment. So, at the end of an RNA sequencing experiment, the goal is to measure or quantify the expression of genes in a sample and maybe compare these expressions across different symbols. So, maybe different conditions, different treatments, etcetera, and see which genes are showing higher expression and which genes are showing lower expression. This gives us tremendous insight into what is actually happening inside the body, inside the cell, etcetera. So, let us look at this example, right? So, we have these three genes, right: gene A, gene B, and gene C, and we have these blue reads mapping these genes, ok? So, we are not talking about exons or isoforms; we are just keeping it very simple here. So, we are seeing these reads that map to gene A, gene B, and gene C in an RNA sequencing experiment, ok? And what we get is raw count data, right? This is something we have seen, right? We have this for many genes; we have this raw count, right, 4, 5, or 1000. So, here is an example, and when you convert these two numbers, the number of reads maps to each gene. For example, for gene A, we have 6 reads mapping to this gene, right? Then for gene B, we have 3 reads, and for gene C, we have 2 reads, right? So, this is something we see. Now, the goal, as I said, is to go to the expression level of genes, ok?


RNA-seq experiment

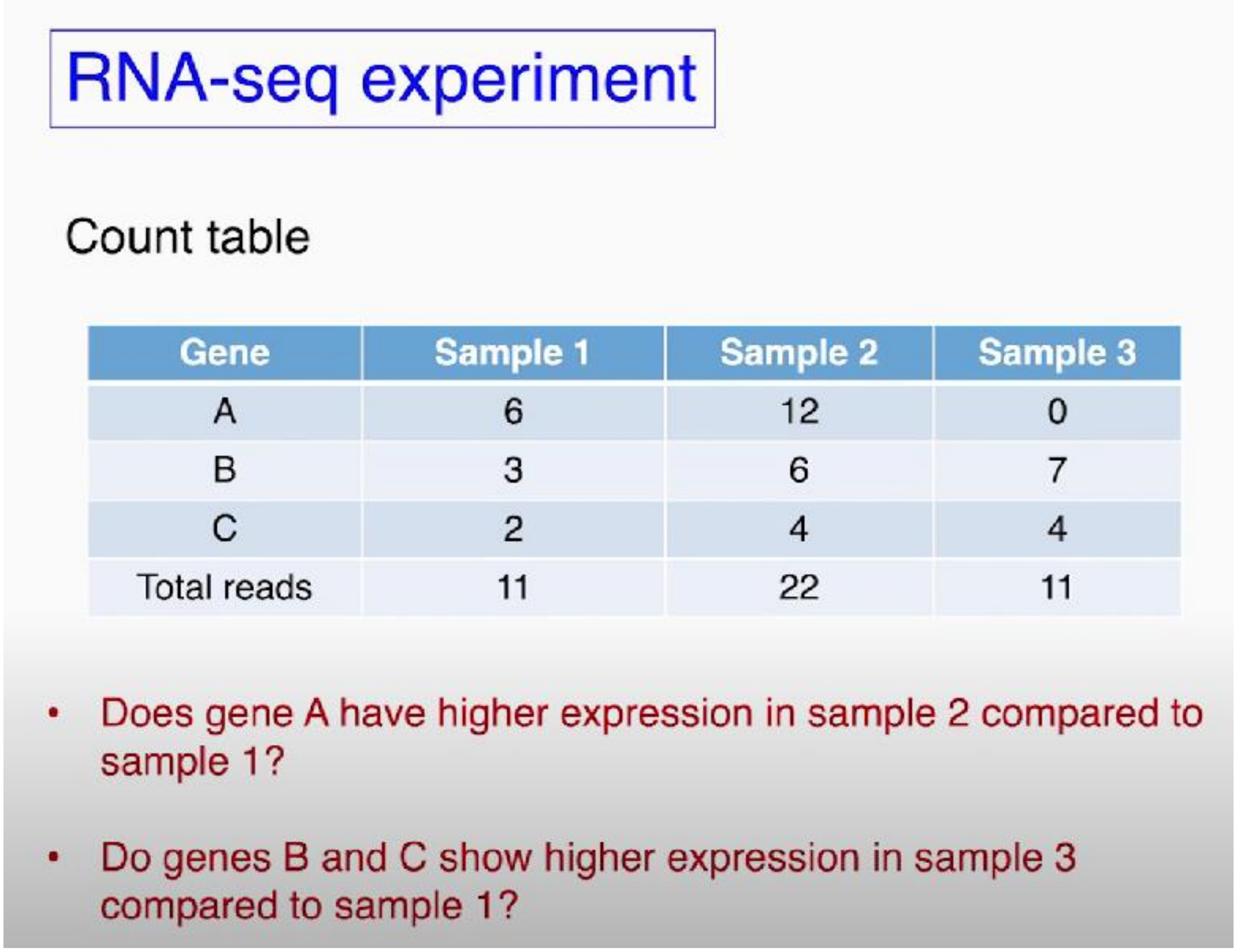
Count table

Gene	Raw read count
A	6
B	3
C	2

Does this mean Gene A has higher expression than Gene B?

So, by looking at this table, can we just say gene A has higher expression than gene B? So, can we just simply say that gene A shows a higher number of reads, right? So, we will talk about this, right? This is not so simple, and we will talk about, right, how we can actually arrive at the expression level from this raw read count data. So, we can actually extend this example further and imagine we now have three samples: sample 1, sample 2, and sample 3. And we do the mapping as before, and we see these counts, right, the number of reads that map to each of these genes in each of these samples, right.

So, we have samples 1, 2, and 3, and we have genes A, B, and C. For sample 1, we see 6, 3, 2, and right reads mapping to genes A, B, and C, respectively. Then we have for sample 2, we have these numbers 12, 6, and 4 and for sample 3, we have 0, 7, and 4, ok. Now, the question is, does this mean gene A has higher expression in sample 2 compared to sample 1 because, as you see in sample 2, there are 12 reads that are mapping to gene A whereas in sample 1, there are 6 reads mapping to gene A? Does this mean this gene A is highly expressed in this sample 2, ok? This is a question we need to answer again because we do a lot of this kind of analysis, where we compare the expression of the same gene across different samples, different conditions, and maybe different treatments.



RNA-seq experiment

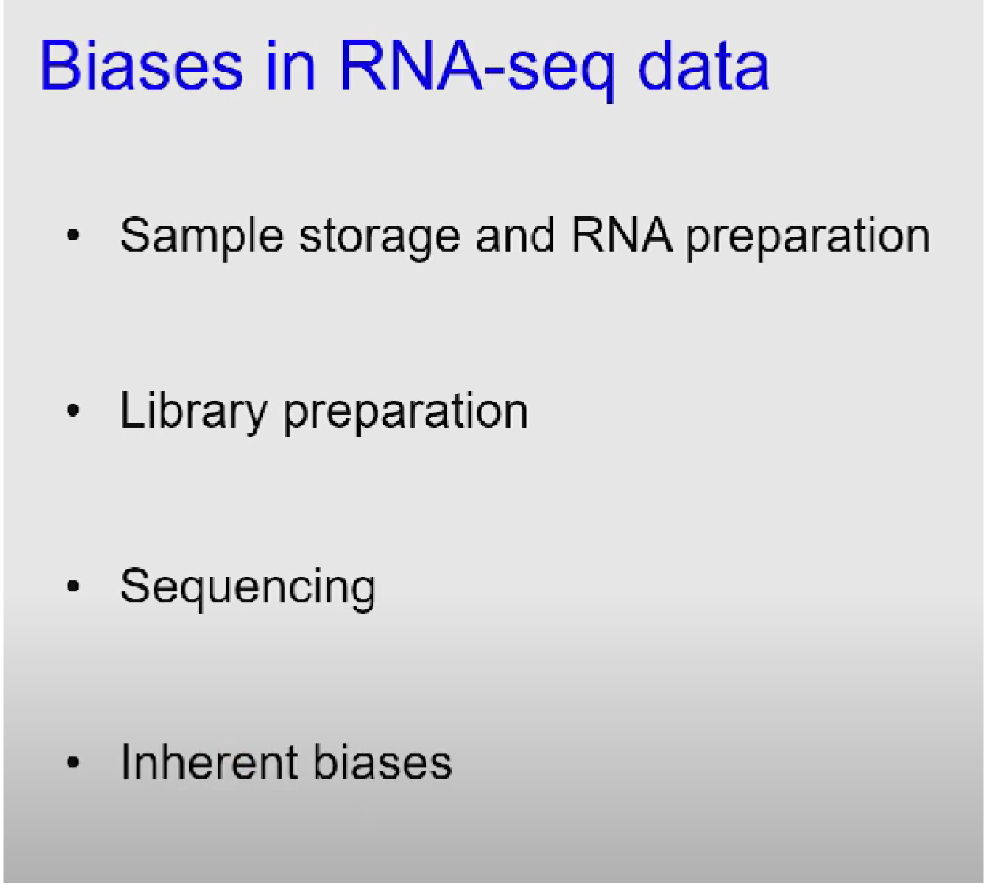
Count table

Gene	Sample 1	Sample 2	Sample 3
A	6	12	0
B	3	6	7
C	2	4	4
Total reads	11	22	11

- Does gene A have higher expression in sample 2 compared to sample 1?
- Do genes B and C show higher expression in sample 3 compared to sample 1?

The second question is, right, if you look at this data again, right, so for genes B and C, does this mean sample 3 has a higher expression of genes B and C compared to sample 1, right? So, you see again that the number of reads that are mapping to gene B in sample 3 is 7, whereas this is 3 in sample 1, and similarly for gene C, we have 4 reads mapping to sample C in sample 3 and 2 reads in sample 1. So, the question is, if you can say B and C are showing higher expression in sample 3, is that so? Again, the answers are not so simple because there are a lot of other factors that we need to take into account, ok, and this is something that we will talk about in this class, ok. So, whenever you do an RNA-Seq experiment, you will always get something called technical variation, and then we have biological variation, ok, and the goal of this discussion on biases and bias correction is to actually minimize this technical variation so that we can study this biological variation, right, because that is where the insight is, right? So, we understand the behavior of the system; we understand how the cell is responding to certain treatment, right, or maybe responding to certain environment, etcetera, ok. So, let us now move into the biases in RNA-Seq data, right.

So, what are the different types of biases that we can see? Okay, and these biases arise due to the nature of transcripts during the library preparation steps and the sequence steps. So, in the next slides, we will talk about these steps where these biases come into play, ok? So, the first step in the RNA-Seq experiment, if you remember the flowchart, is the sample preparation, right? So, we have sample storage sometimes, right? If you are working with tissue data, you have to collect these tissues, and maybe you have to store them in certain conditions. So, and then you have the RNA preparation path, right? So, if you are isolating RNA, right, you will follow certain protocols that can affect the quality of these RNA-Seq data, right, and which kind of read data you will get. Similarly, we have the library preparation step, which can also introduce some biases in the data. You have the sequencing step; again, depending on the sequencing platform, you have amplification on some platforms, etcetera. Those can affect the quality or maybe introduce some biases in the data.

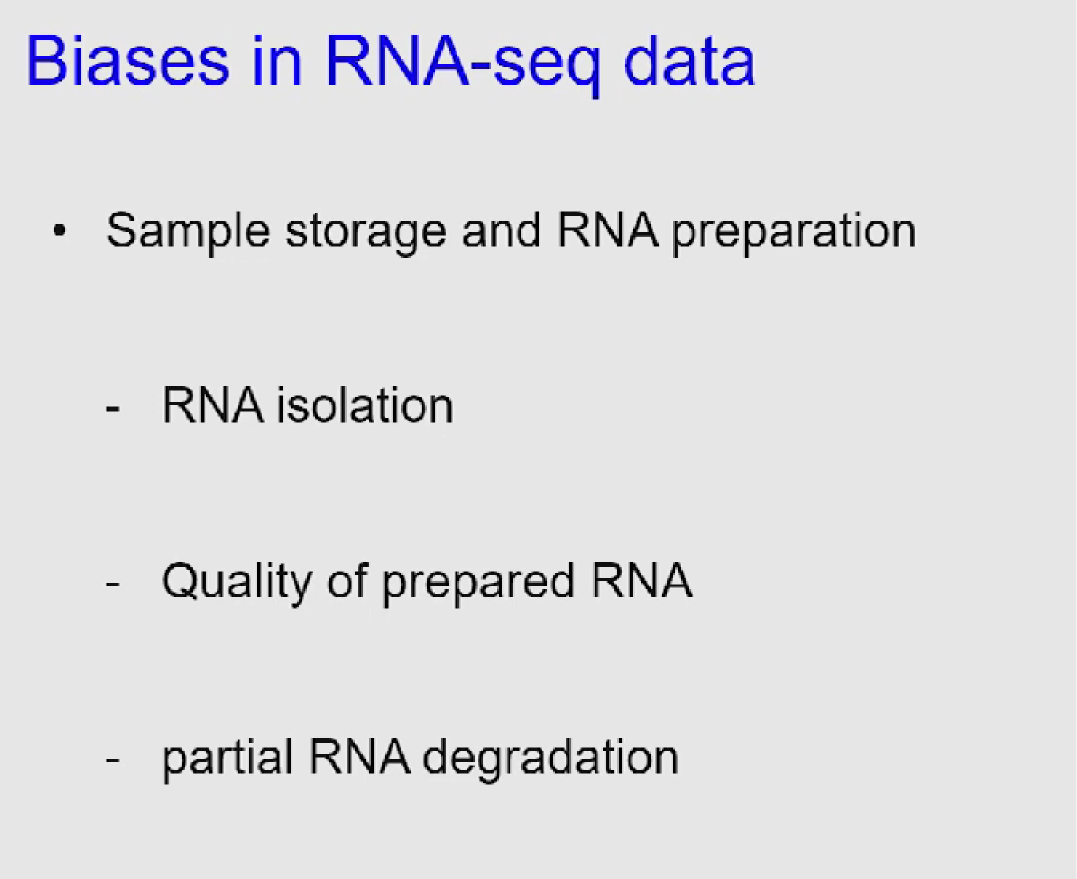


Biases in RNA-seq data

- Sample storage and RNA preparation
- Library preparation
- Sequencing
- Inherent biases

And of course, there are some inherent biases in the RNA-Seq experiment, which we will talk about at the end of this class. So, let us now take each of these points and see how they actually introduce biases in the data. So, the first step is sample storage and RNA preparation, as we have mentioned. So, we have different types of RNA isolation protocols, and this might affect the RNA-Seq data, right? So, you can have different types of biases and different types of RNA processes that you are utilizing that actually will determine, right? So, again, the quality of the prepared RNA is very important here, as you can probably understand, because there are a lot of these RNA molecules around. So, you can get maybe partial degradation of some molecules, and maybe some RNA molecules are degraded, and this can affect or bias your data, right? So, you can imagine, right, if out of the full transcript, some of the transcripts are degraded by these RNAs, you are biasing your data. So, you will see that maybe some genes are poorly expressed, and maybe in the actual scenario, these transcripts have been degraded by the presence of RNA. So, this is very important, right? The quality of the prepared RNA is a very important step.

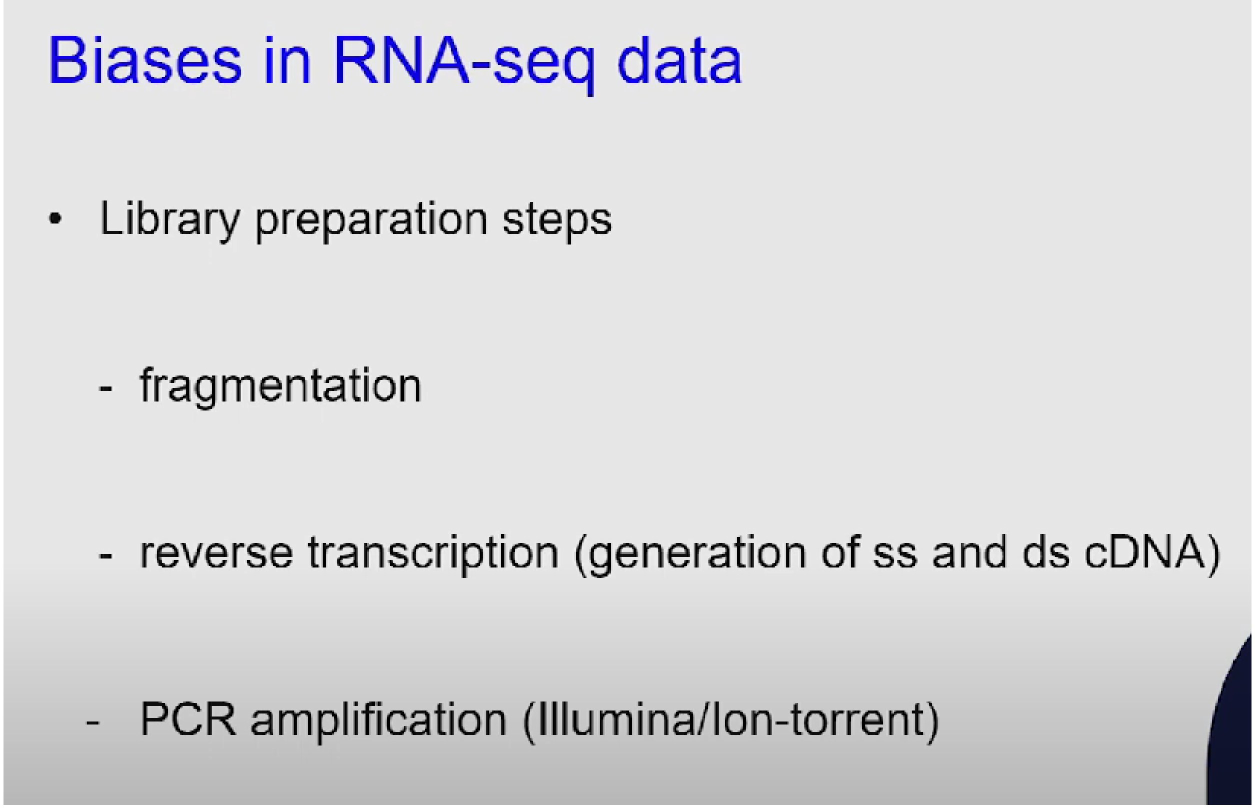
And as I mentioned, you can see this partial RNA degradation, and if you have this partial RNA degradation, that will bias your data, right? So, again, this is an experimental step that can ultimately affect the data that you see after next-generation sequencing, ok? So, is there any way we can actually deal with this?



Biases in RNA-seq data

- Sample storage and RNA preparation
  - RNA isolation
  - Quality of prepared RNA
  - partial RNA degradation

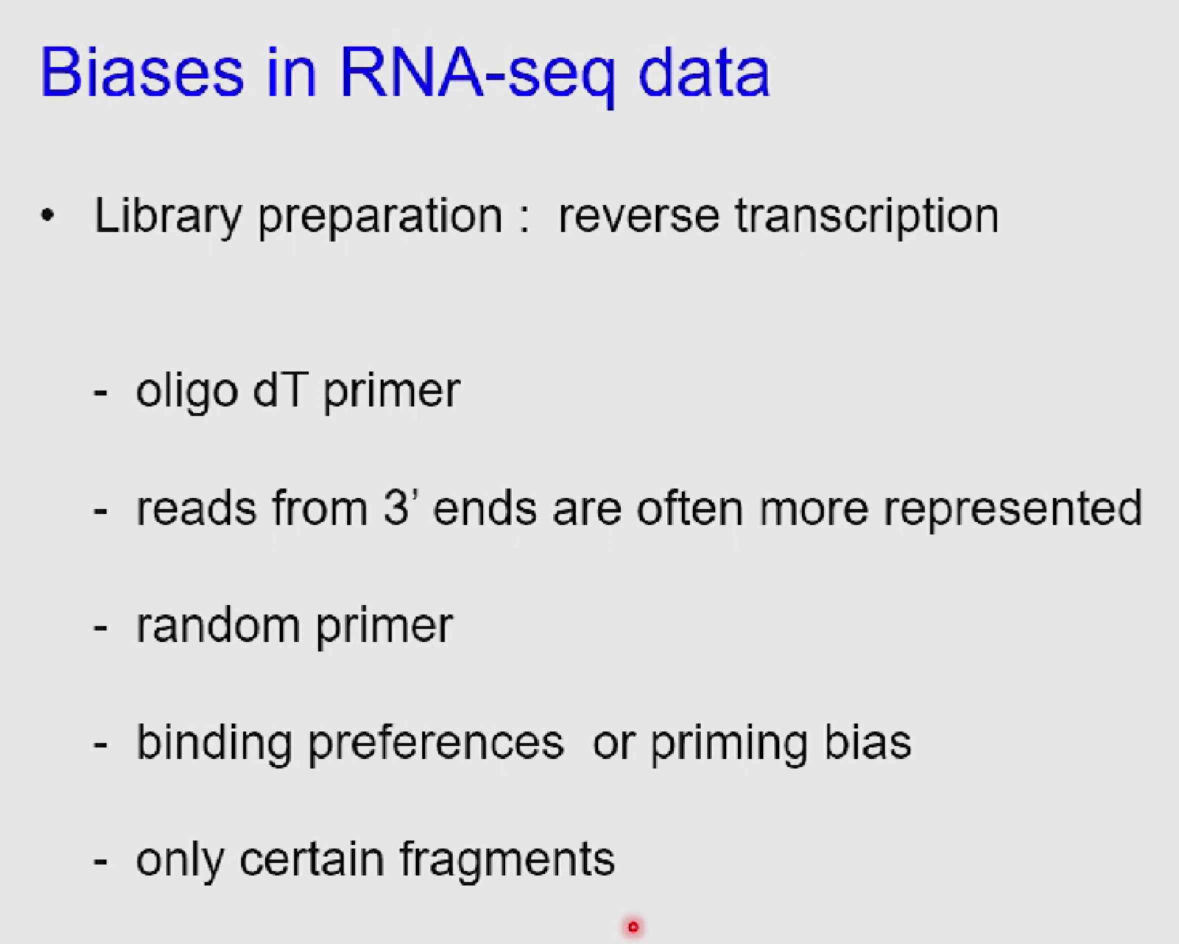
So, this is something that is very important, and we will come to that towards the end of this class. Now, moving on to the library preparation steps, So, there are different steps when you prepare an RNA-seq library; we have talked about them, right? So, one is fragmentation, right? So, we have to fragment long transcripts with different steps, and you have the reverse transcription, right?



Biases in RNA-seq data

- Library preparation steps
  - fragmentation
  - reverse transcription (generation of ss and ds cDNA)
  - PCR amplification (Illumina/Ion-torrent)

We have talked about this. So, this is the generation of cDNA, double-stranded cDNA, and then sometimes you also have PCR amplification, right? If you are going into Illumina or ion torrent platforms, you need to have on-grid amplification or cluster generation, right? So, bridge amplification. So, that is, it can also introduce bias in the RNA-seq data. So, let us talk about fragmentation and how it can actually affect the RNA-seq data. So, you usually have random fragmentation, right? So, this can happen at any place in the transcript, and you can do the fragmentation either at the RNA level, right? You can fragment the RNA, or you can fragment the cDNA, right? And what happens is that after this fragmentation, because this is a random process, you generate fragments of different sizes, and before you go for sequencing, you have to do something called a size selection, ok. Again, if this fragmentation process is biased towards certain sites, right, you can see only certain fragments are getting selected at the size selection step and long fragments are excluded from sequencing, ok. And this has been observed, right, in many experiments, that these starting points of fragments actually coincide quite often. This is not supposed to be the case, right? So, if fragmentation is random, then you should see a lot of these different starting points for fragments, right? But what researchers have observed is that these fragments and the starting points often coincide, suggesting that this is not a random process. So, we think this is a random fragmentation process; this is not okay. And again, this introduces bias in the data, which we need to account for. So, we would have to do something about this at the bioinformatic step, right? We can probably change these protocols too much; maybe there are some sort of improvements that you can make, but then we have to have these bias correction steps in the bioinformatics pipeline. Then we have the reverse transcription, right? So, for many library preparation steps, we need to do the reverse transcription, except probably the direct RNA sequencing that we have talked about. So, in reverse transcription, the earlier researchers used this oligo-dt primer where they are looking at the poly A tail, right. So, these primers will bind at the poly A tail. So, what we will see when you have this kind of oligo-dt primer you are using is that you will see only certain parts, right? From the 3 prime ends of the RNA, they will be amplified, and maybe you will see a lot of overrepresentation of the 3 prime ends of this RNA molecules, right? So, you will see more reads there and you will see fewer reads from the 5 prime ends, ok?



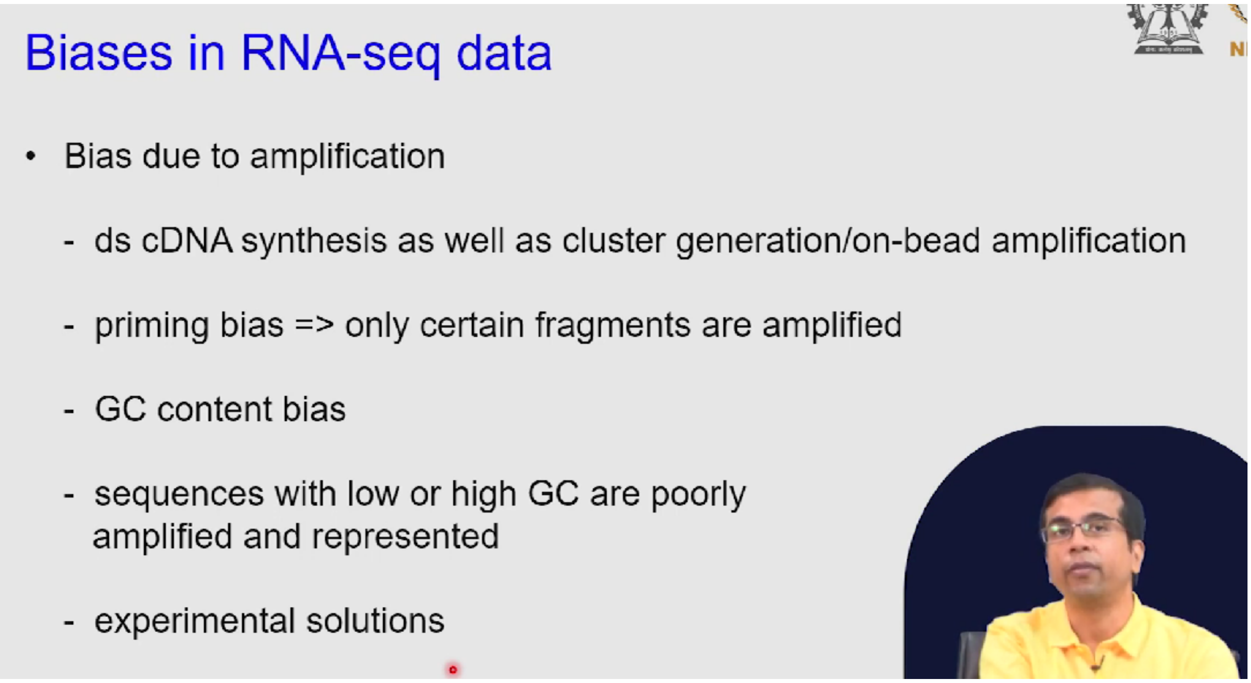
Biases in RNA-seq data

- Library preparation : reverse transcription
  - oligo dT primer
  - reads from 3' ends are often more represented
  - random primer
  - binding preferences or priming bias
  - only certain fragments

So, this can happen, and again, as we have talked about, this means we have non-uniform coverage across the transcript. Now, what researchers suggested is that maybe we can use random primers instead of this oligo-dt, right? So, this should solve the purpose, solve this problem of this bias towards certain regions, right. But again, we have binding preferences, right? These are usually random primers. They will again have binding preferences for certain regions, which we call priming bias, and this can again lead to the generation of only certain fragments being overrepresented in the read data. So, again, you can try different ways, but again, these biases will be in the data.

You can also have bias due to the amplification step, right? So, this is amplification in the cDNA synthesis step or in cluster generation on-grid amplification, right? So, you can generate this bias. Now, we have talked about priming bias. So, it means we are getting only certain fragments that will be amplified.

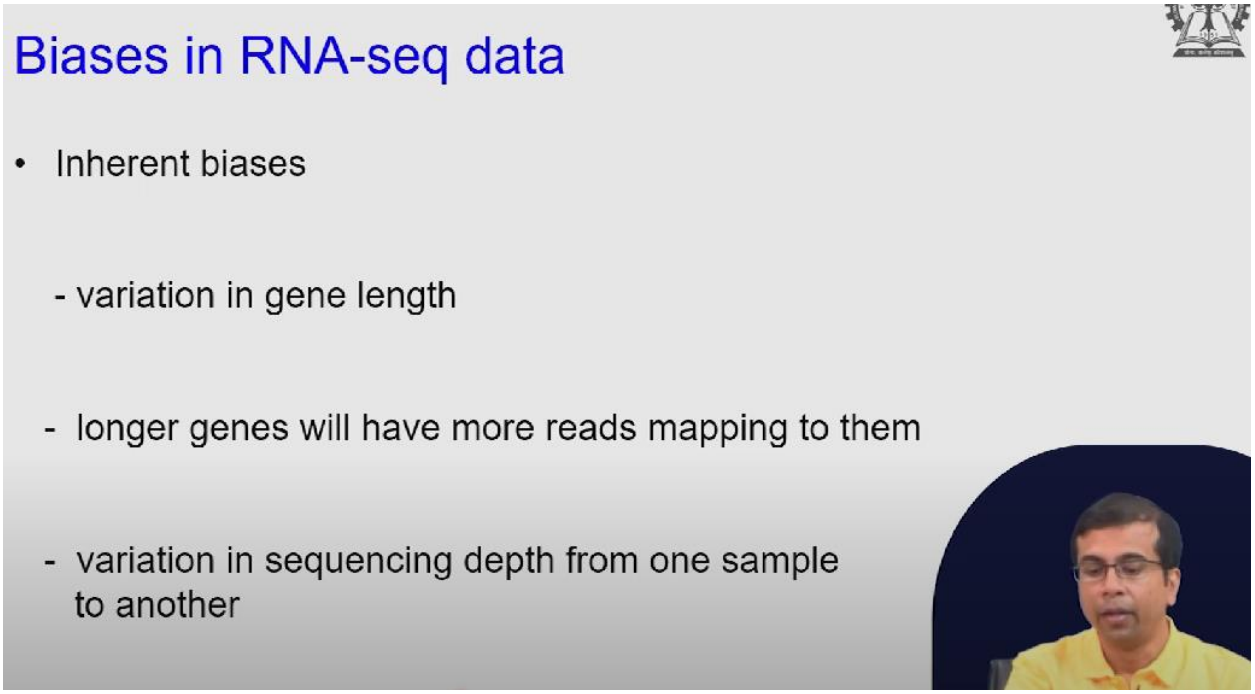
So, these will be overrepresented in the library. In addition, there is something called a GC content bias, ok? So, what is this GC content bias? Is that what we have observed or researchers have observed is that sequences that have low or high GC are usually poorly amplified and poorly represented in the RNA-seq library, ok? So, again, this is something to do with the amplification process, and you can have certain modifications in the amplification step, right? This is an experimental modification. You can add certain solutions that will help in better amplification of these lower high-GC regions, right? So, you get a more or less uniform representation of all these regions in the RNA-seq library. So, you have some experimental solutions, but again, you do not know how effective this solution is until you see the read data, and from the read data, you can identify whether this kind of bias is still there.



Biases in RNA-seq data

- Bias due to amplification
  - ds cDNA synthesis as well as cluster generation/on-bead amplification
  - priming bias => only certain fragments are amplified
  - GC content bias
  - sequences with low or high GC are poorly amplified and represented
  - experimental solutions

And this is something that is a very important part of all this analysis. And finally, what we have talked about is that there are inherent biases in the data, right? So, one of the first things is the variation in gene length, ok? So, you have, let us say, 10,000 genes, right? They vary tremendously in terms of the size, right, of the transcript. So, some transcripts may be 1 kB in length, some transcripts may be only 200 base pairs, and some transcripts may be 5 kB. And this also leads to bias in the data. So, we will talk about this, right? So, what happens is that if you have a longer transcript, there is a high chance that you will get more fragments, and you will have more reads coming from this transcript in the RNA-Seq library, because simply because this is a longer transcript, you have more fragments generated, and that will be sequenced more than a transcript that is smaller in length. So, this is what actually would lead to bias in the data, ok? So, you will find you have longer genes and more reads mapping to them, usually. If you are comparing this between a long gene and a short gene, and they are expressed at the same level you will usually see longer genes have more reads compared to the short gene, ok?



Biases in RNA-seq data

- Inherent biases

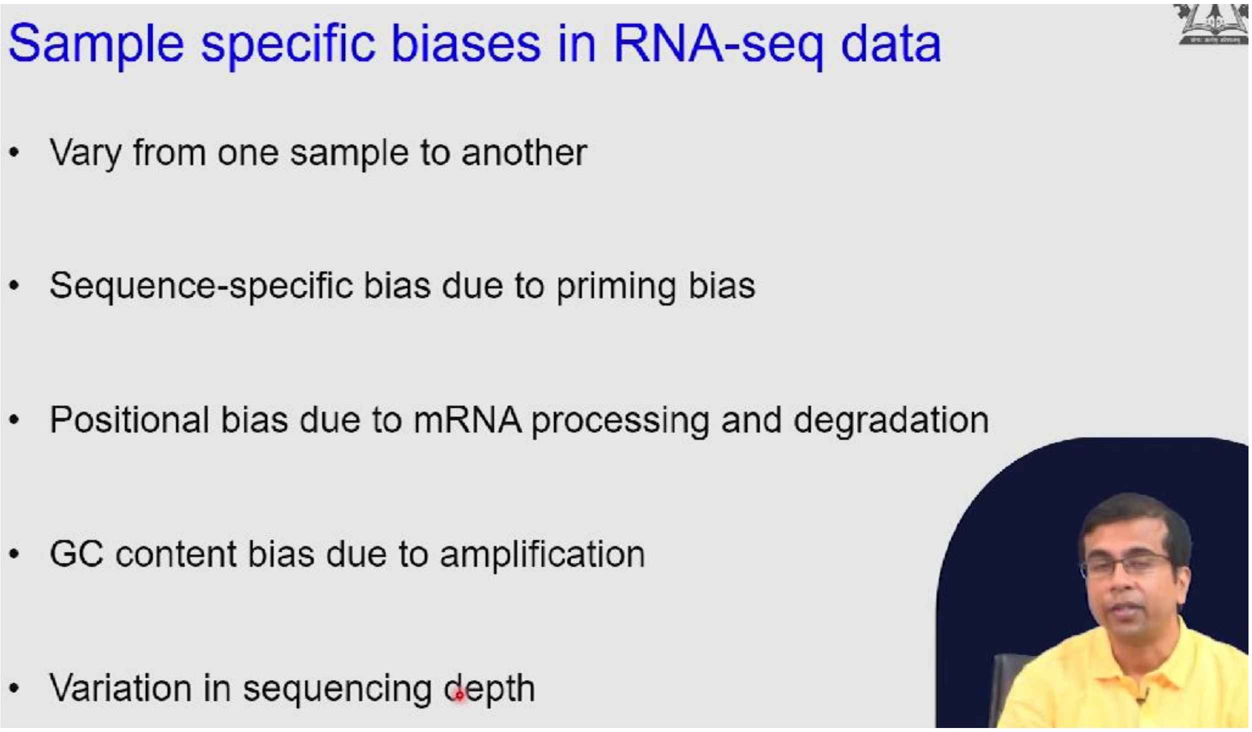
  - variation in gene length

  - longer genes will have more reads mapping to them

  - variation in sequencing depth from one sample to another

So, this is a bias that is present in the data. In addition, you have something called variation in sequencing depth from one sample to another, right? So, if you sequence two samples, maybe in one sample you get 1 million reads in total, and in another sample you will probably get maybe 1.5 million reads, right? So, again, you cannot control how many reads exactly you will get from each sample, right? So, you will see this variation, and again, you need to account for this variation, ok? So, we have talked about these different types of biases in the data that you can see. Some are sample-specific biases, right? So, when I say sample-specific biases, these are present in each sample and unique, right? So, they are unique for a sample, right? This is not like you have processed these eight samples, and you see the same bias across this sample. That will actually be easier for us to deal with because if you see this kind of uniform bias across this data set, you can filter out these biases quite easily, or maybe if you are doing comparative analysis between samples, you may not have to counter or account for these kinds of biases. But what we see is that these biases are sample-specific. And this means we have to correct for these biases at the sample level, ok? And this, as I mentioned, is unique to each sample.

So, they vary from one sample to another. You have sequence-specific bias. This is due to priming bias, right? So, again, if you are using random primers, for example, reverse transcription, you will see that only certain regions will bind these random primers. They are amplified much more efficiently, right? Again, this is something we cannot really do anything about because if you are using these random primers, And this will lead to something called sequence-specific bias, right? So, again, these regions that bind to these primers will be overrepresented in the RNA-Seq library, and the other regions will be poorly represented. We have also talked about something called positional bias, right? So, this is again something to do with mRNA processing and degradation, right? So, if you are processing this mRNA, if you are using this oligo dt and maybe some ends, let us say one end of the RNA is degraded, right? So, other than that, most of the reads will come from a specific position within the transcript, ok? And then we also talked about this GC content bias, right? So, this arises due to the amplification step. So, some regions that have a neutral GC content of close to 50 percent are actually amplified much more efficiently than the regions that have a very low or very high GC content. And we talked about some of the experimental improvements that you can make to actually reduce this GC content bias.

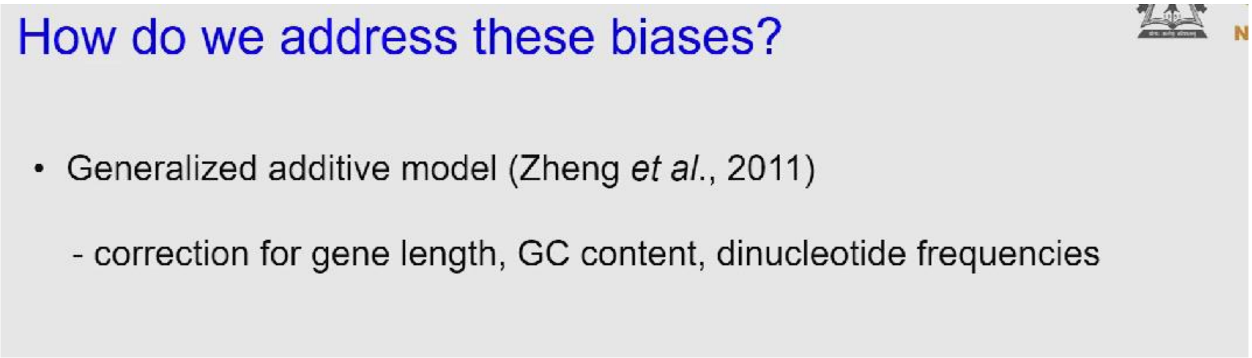


Sample specific biases in RNA-seq data

- Vary from one sample to another
- Sequence-specific bias due to priming bias
- Positional bias due to mRNA processing and degradation
- GC content bias due to amplification
- Variation in sequencing depth

And then finally, we also have this variation in the sequence step. This again varies from one sample to another. So, one thing you should note is that we have not included gene length bias here because this will not vary from one sample to another. This is the same across all samples or likely to be the same.

Now, the question is, how do we address these biases? So, we have understood these different types of sample-specific biases, right? And we need to remove these biases from the data before we actually further process the data, right? So, these are technical variations; they are introduced because of the experimental process, the RNA preparation process, or the library preparation process, right? These are not actual biological variations, ok? So, we want to minimize or remove these technical variations so that we can study the biological variation in the data. Now, how do we address this? So, we can learn about these biases from technical replicates of the same sample. So, we use two terms: one is technical replicates, and the other is biological replicates, right? So, technical replicates means you are doing the same experiment, right, with the same sample multiple times, ok? And this will help us understand, right? If you see this variation in read data across these technical replicates, you can then say, "Okay, this is where these are the different types of biases that we see and we can model these kinds of biases." And we can take different approaches for bias correction, and this has been proposed, right? Many methods have been proposed, and we can discuss some of these methods in the next few slides, ok? So, one of the methods that have been proposed relies on something called a generalized additive model.

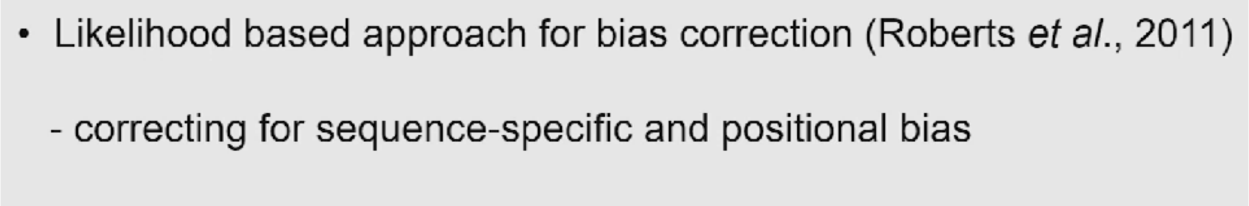


How do we address these biases?

- Generalized additive model (Zheng *et al.*, 2011)

  - correction for gene length, GC content, dinucleotide frequencies

This is the paper; again, the reference is given at the end, and it actually does a correction for gene length, GC content, and DNA nucleotide frequencies. So, we have talked about these biases because of gene length, right? So, this again varies from one gene to another. So, this has to be taken into account. We have GC content bias, and we again discuss how this arises because of the amplification bias for specific GC contents. Then you have dinucleotide frequencies; again, this will arise if you have sequence-specific biases, right? So, certain sequences are preferred by the primers that are binding, right? So, you can then see these dinucleotide frequencies changing, right? Some dinucleotides will be much more abundant than the others, ok? So, this method can actually take into account all these biases and correct for them. You also have something called a likelihood-based approach to bias correction.

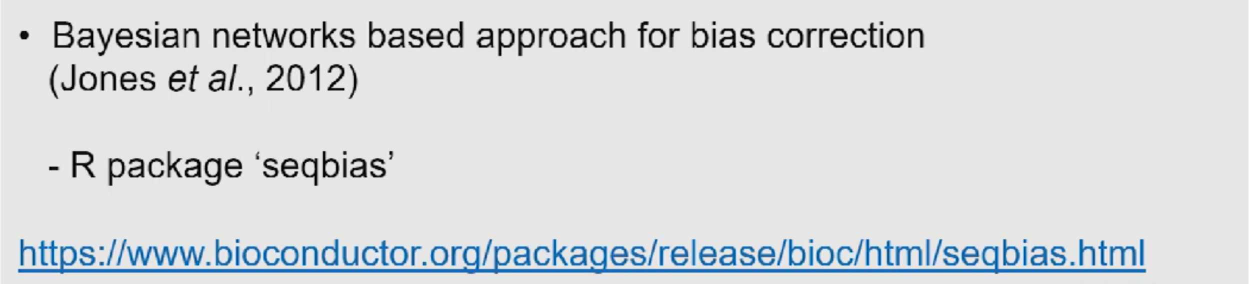


- Likelihood based approach for bias correction (Roberts *et al*., 2011)

  - correcting for sequence-specific and positional bias

Again, we are not going into the details of all these methods, ok? So, what does it do? It is actually looking at this expression variation as well as the bias, right? So, it kind of develops models, right, that actually kind of generate estimates for this bias, ok? And it can correct for sequence-specific and positional biases. There are other methods. For example, this is a method that has been proposed using a Bayesian network-based approach.

Again, we are not going into the statistics. If you are interested, you can take a look. And you can see this package that has been developed. So, you can take this tool, right, in R, and you can actually estimate these biases in your own data set, ok? So, again, we will introduce R when you go into the hands-on, and we will see how we can utilize these different packages for application to our own data sets, ok? So, this is a method that actually can learn the type of bias that is present in the data and can correct for those biases.



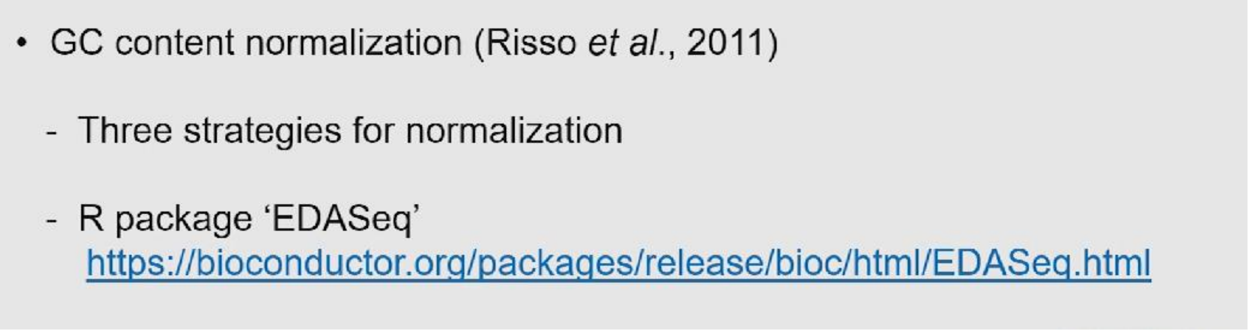
- Bayesian networks based approach for bias correction (Jones *et al*., 2012)

  - R package ‘seqbias’

https://www.bioconductor.org/packages/release/bioc/html/seqbias.html

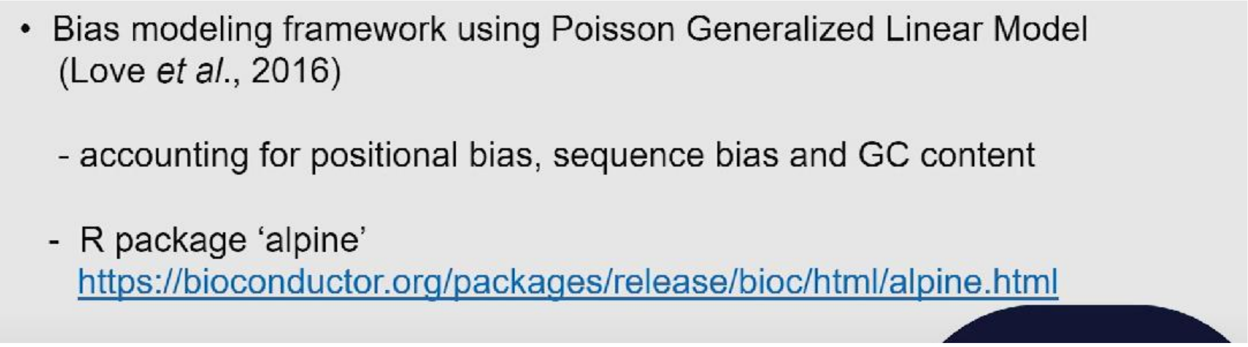
Now, there are specific methods that have been proposed, for example, for GC content normalization. So, we have learned about this GC content bias, right? This arises because of the difference in amplification efficiency between different fragments with different GC content. So, this actually has been proposed in this paper that is mentioned here, and they have suggested three strategies for normalization.

So, again, you can have a look into the details. We are not going to discuss all the strategies, the different types of methods, etcetera. And again, there is an R package that they have developed called EDASeq. And again, here is a link you can see, download, and install if you have some knowledge of R. Otherwise, do not worry; we will actually discuss how we can install these R packages and how we actually use them to correct for these biases in our own data sets.



- GC content normalization (Risso *et al*., 2011)
  - Three strategies for normalization
  - R package 'EDASeq' https://bioconductor.org/packages/release/bioc/html/EDASeq.html

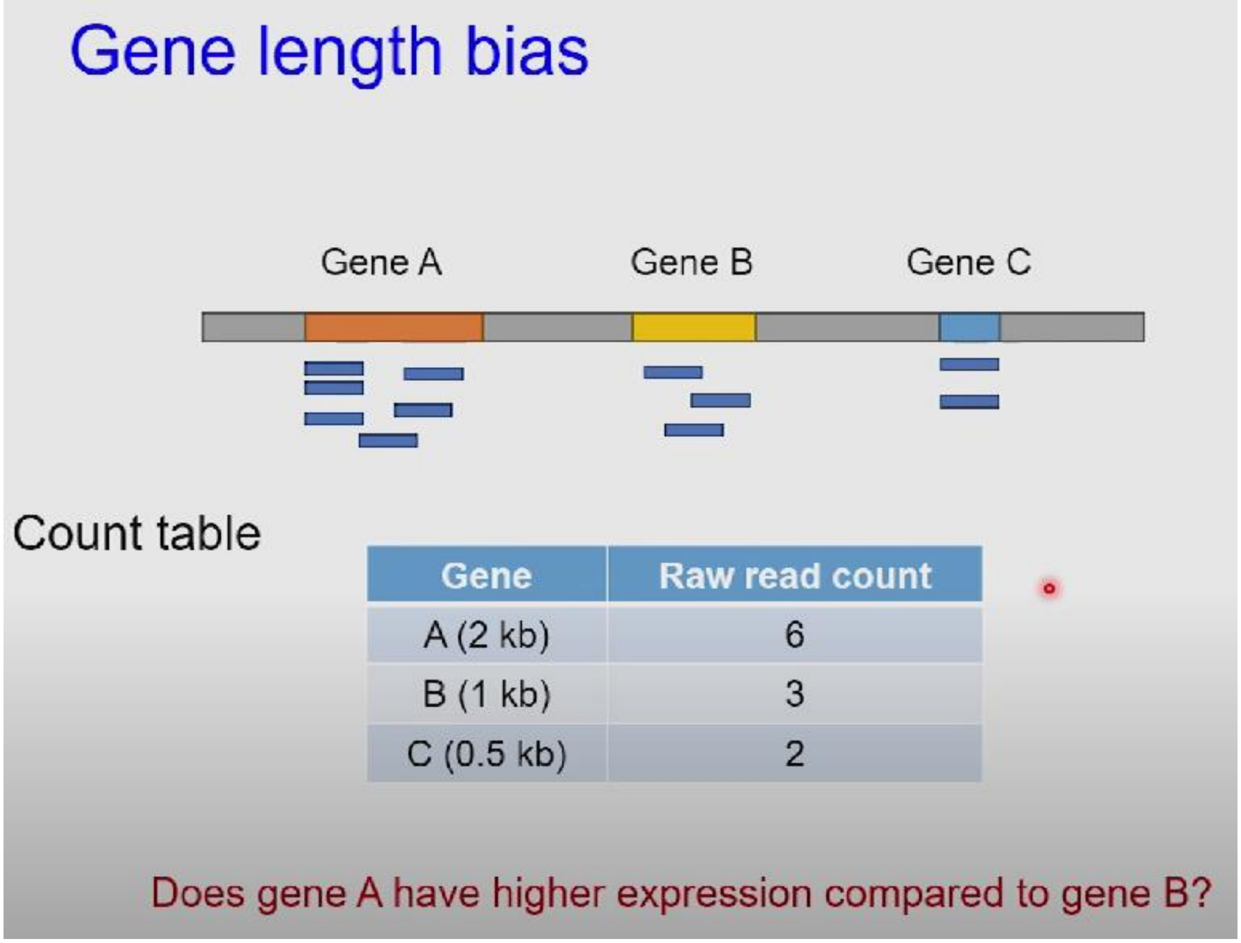
So, we will do that in the hands-on part. There is another method that actually relies on a bias modeling framework and utilizes something called the Poisson Generalized Linear Model. Again, these technical terms are not so important. If you are interested in the actual statistics and methods, please have a look at the original paper. We are not going to go into the details of this.



- Bias modeling framework using Poisson Generalized Linear Model (Love *et al*., 2016)
  - accounting for positional bias, sequence bias and GC content
  - R package 'alpine' https://bioconductor.org/packages/release/bioc/html/alpine.html

This method can account for positional bias, sequence bias, and GC content. So, these are the three major biases that we have seen that we have discussed previously. And they have also developed a package called alpine, which is again an R package that you can download and apply to your own data. Now, we will see that most of these packages for transcriptome analysis were actually developed as R packages, and we would have to actually use R to analyze these data sets. So, we will do that in the hands-on part. So, one of the things that I have mentioned and one of the points I mentioned is that there is a gene length bias right, and what we originally assumed was that the expression level of a gene right is proportional to the number of reads that map to the gene right.

So, this is a very naive assumption, and again, this will be affected by this gene length bias. So, this is the assumption, right? So, the number of reads mapped to a gene is proportional to the expression, but what we see is that this is not the case, and this number actually also depends on gene length, ok? And as we have discussed, if you have a longer transcript, you have more fragments inherited that can be sequenced in the library. So, these fragments will be overrepresented in the library, and they will appear to be more highly expressed, even though they are not okay. So, if you compare a long gene and a short gene and compare if they are expressed at the same level, you will see that the longer gene will show a higher number of reads mapped to them because they simply have more fragments generated from them. So, coming back to the same example that I gave you, So, again, you have these three genes gene A, gene B, and gene C, and we have the countable. Now, what will we do since we know about this gene length bias? I will also mention the gene lengths right here, ok? So, here you see now that gene A is of length 2 k B, gene B is 1 k B, and gene C is 0.5 KB.



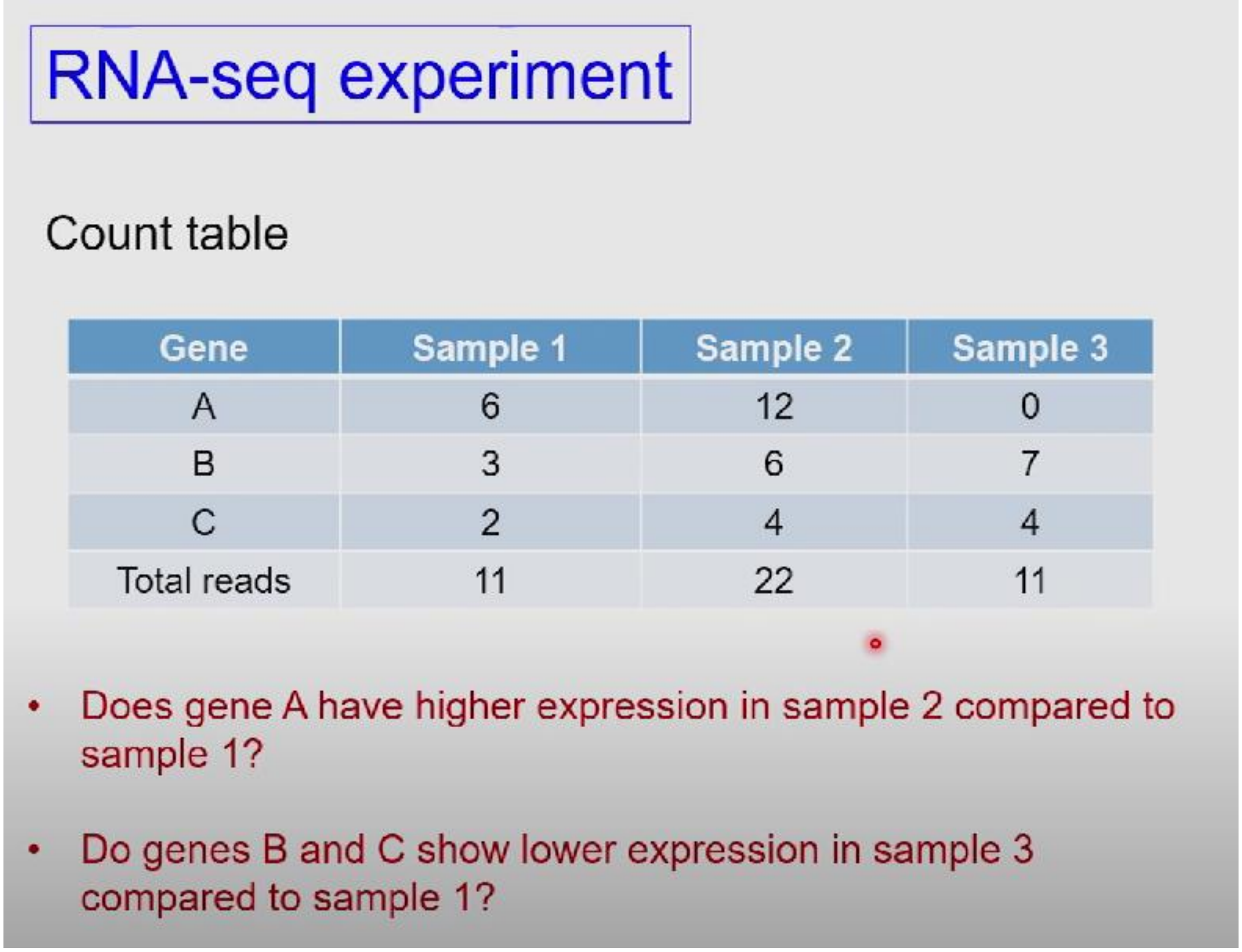
Gene length bias

Count table

Gene	Raw read count
A (2 kb)	6
B (1 kb)	3
C (0.5 kb)	2

Does gene A have higher expression compared to gene B?

Now, once we know this right and we know about gene length bias, So, can we answer this question now? Does gene A have higher expression compared to gene B? The answer is no. So, because gene A is longer, it is actually twice right. So, the length is 2 kB compared to 1 kB of B, and it should have around 2 times more reads mapped to it, ok? So, this is something again that now tells us this is not true, right? So, gene A and gene B probably have very similar expression levels, ok? So, you can now see that when we are accounting for these biases in these questions the answers to the questions now become clear, and we are actually getting the right answers. We are not making wrong influences, and we can actually extend this to this other example that we have taken. We are comparing these expression levels across samples 1, 2, and 3, and again, we have this question: does gene A have a higher expression level in sample 2 compared to sample 1, and do genes B and C show lower expression in sample 3 compared to sample 1? Now, what you have done compared to the earlier table is that I have now added this total number of reads here. So, this is the total sequencing depth, and as we have discussed this depth can vary from one sample to another. So, this is something we again have to take into account, and you can probably imagine that if you have a higher total number of reads in a sample, you are likely to see more reads mapping to each gene in that sample. So, if you take this example here, if you have 22 reads in this sample 2, compared to 11 reads in sample 1 and 11 reads in sample 3, So, naturally, you will expect more reads mapping to genes A, B C right in this sample 2 compared to sample 1. So, what you usually can do is we can actually account for this total number of reads, and we can look at the fraction of reads that are mapping to the genes as compared to the total number of reads that are mapping to the genes, and this is something that we will discuss next.



RNA-seq experiment

Count table

Gene	Sample 1	Sample 2	Sample 3
A	6	12	0
B	3	6	7
C	2	4	4
Total reads	11	22	11

- Does gene A have higher expression in sample 2 compared to sample 1?
- Do genes B and C show lower expression in sample 3 compared to sample 1?

For the second question here, do genes B and C show lower expression in sample 3 compared to sample 1? you have to be a bit careful, right? So, one of the things you probably notice is that gene A becomes 0, right. So, it means clearly that gene A's expression has dropped in sample 3. So, what it means is that if you are sequencing up to the same depth, only B and C will get all the reads because gene A is not expressed now, ok? So, only genes B and C are expressed. So, all the reads will come from genes B and C, but this does not mean that genes B and C are showing higher expression because gene A's expression has dropped. So, we are not getting any reads for gene B and gene C; they are showing a higher reach map to them, ok? And this kind of problem, right? This kind of issue will arise, and this kind of bias will arise. This is an inherent bias in the RNA-seq experiment, and you will see similar kinds of problems. If gene A is very highly expressed it will appear that in genes B and C, their expression has diminished. It is because the number of reads that you are getting remains the same; only gene A will take up a much bigger fraction of the total number of reads, but this will reduce the fraction of reads for B and C, and they will appear to be downregulated or showing lower expression. So, this is something that we again have to be careful about, and this is something we have to deal with when we are analyzing the data. So, the technical term that we use is called data normalization. This is the process that actually tries to minimize these technical errors, and here is the pipeline again for you. So, we have talked about all these steps before quality control mapping assembly and quantification, and this is where we are now normalizing, ok? And this is the final step before we actually go into the data analysis or actually answer the biological question that we are interested in. o, in data normalization, the goal is to minimize technical variation without affecting biological variation, and we have to do something called sample-specific normalization, or sample normalization. So, we have talked about some of these bias correction methods; these are actually within sample normalization, and then you also have between sample normalization within a data set. So, if you are working with multiple samples, as in the example that you have taken, and we want to compare across samples, then in that case we would have to do sample normalization, but these are part of a single data set, and sometimes we are working with multiple data sets, and in that case we have to do something called a batch correction. So, again, this variation from one set of experiments to another can be seen, and we have to correct for those. So, within sample normalization we have talked about bias correction methods, for example, positional bias, sequence-specific bias, and GC content bias.

We have also mentioned gene length bias as well as sequencing depth. So, we will discuss in much more detail the data normalization method in the next few classes. Here are the references for this class, and to finally conclude, we have talked about these different types of biases in the data, and these can generate technical variations in the data that we need to account for. We have talked about sample-specific biases, which include positional bias, sequence-specific bias, and GC content bias. We have also discussed different bias correction methods, which can actually account for these different biases. So, we get improved transcript abundance estimates, and we have talked about some of these tools that we can use for this process. We have also introduced a term called normalization, which is actually a method that is aimed at minimizing technical variation. And we have again discussed this within the sample and between sample normalization and what the need is, and we have also discussed batch correction.

So, in the next class, we will talk about these different types of normalization in much more detail.

Thank you.