Next Generation Sequencing Technologies: Data Analysis and Applications

Sequencing coverage, Quality score and Experiment design

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. We have so far discussed all the sequencing technologies that we use today, and in this class we will be talking about some very important technical terms that we will be using throughout this course. So, these two terms are given here; the first one is called sequencing coverage. So, this is very important for designing any next-generation sequencing experiment, and we will be talking about quality score. So, this quality score, as we will see, is very important for interpreting the data or ensuring that the data that we are using for interpretation is good.

And finally, we will have some discussion on experiment design. So, maybe at some point you will have to design an NGS experiment, right? You want to use these NGS methods for your project or for some work, right? And you might be wondering which platform I should use, what the read length should be, how many reads I need, and what kind of sequencing method I should adopt.

So, these questions, of course, require careful examination of different parameters. So, at the end, we will discuss the design of this experiment. So, these are the keyword coverage, quality, and design aspects in this class, ok? So, just to start with a brief overview of this NGS data analysis method, ok? So, here is the flow chart, ok?

So, in the first step, we have this data from a sequencer, right? Whichever sequencer you use, we have this data. This is quite a significant amount of data, usually, right? These data sets are huge, and again, the data format could be different, ok? So, we will talk about this data format in the next class because these sequencers give out data in different formats.

So, you will have to deal with these different data formats, ok? Now the first step that we have in this data analysis process is the data quality check, or QC. We want to ensure that the data that we have is of good quality. Otherwise, if we do all the downstream analysis without doing this quality check, the conclusions that we would get might be erroneous or might be wrong. And this is something we want to prevent, right?

So, at the onset, we want to do this data quality check, or QC. Now you can take different parts after you have done the quality check and you see that you are satisfied that the data is good. Then you can take different parts for different types of analysis, ok? So briefly, one part that we have is the assembly part. Okay, read assembly. So this is true for new reference genomes, right?

So if you are sequencing a new genome or new organism, for which there is no genome data available, the first thing we do is do an assembly, read the assembly, and get the new reference genome. In other cases where you have a reference genome already, what we do is something called read mapping, ok? And this read mapping happens against a reference sequence, ok? Again, these two steps are completely different, right? This read mapping and read assembly are quite different, and they have their own independent algorithms for doing these steps.

We will discuss this in later classes. This reads about the assembly process, the algorithms, the tools, etcetera that are out there. Similarly, for read mapping, we will talk about the algorithms and the tools. Now once you have done read mapping, again, this read mapping is required for a lot of analysis and a lot of different types of analysis. For example, if you are interested in genome analysis and, for example, are looking for genetic variations or structural variations, such as single-link polymorphisms, you will do this genome analysis, and the first requirement is the read mapping.

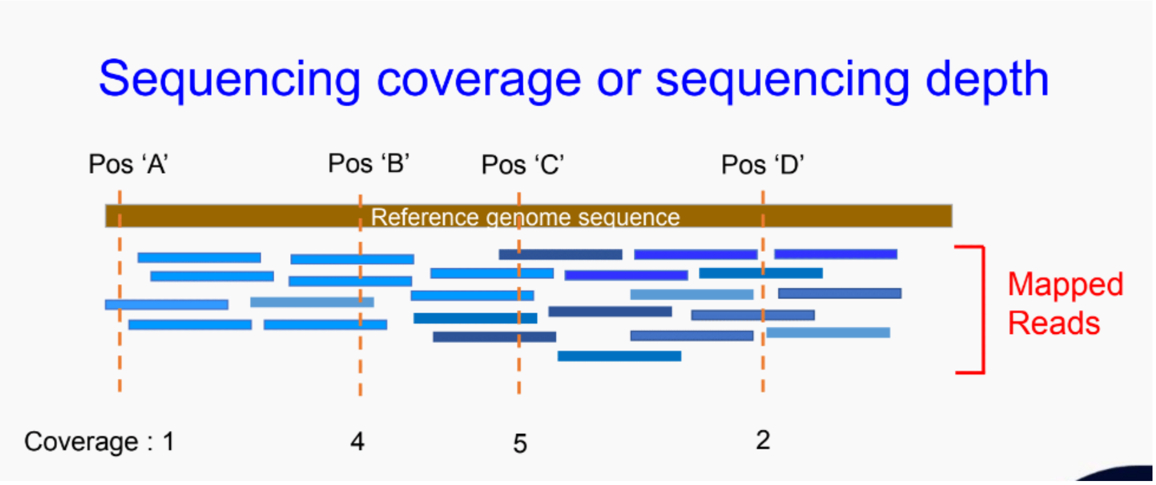
You can also be interested in measuring the expression level of mRNAs, right, or genes, right, and we will do something called transcriptome analysis, and the first step is again read mapping, ok? And then finally, for epigenomic studies, you also need to do read mapping, ok? So we will talk about these steps in much more detail, ok? For these genome analysis steps, identification steps, and structural variations, we will have a dedicated section on the transcriptome analysis part. We will also talk about the epigenomic studies. So the assembly problem can be summarised in this very simple cartoon, right?

So what you have, what you get is the reads, right, from the sequencer here on top, ok? So the reads are shown in different shapes of blue, and what you want to do from these reads is build a single continuous sequence, right? That is the goal, ok?, and this is what we call the reference genome sequence, right? For example, the human reference genome, right, or the mouse reference genome, right So we have these DNA fragments, right? We isolate cells, we have these DNA fragments, we sequence using a sequencing platform, we get these reads, and finally, the ultimate goal is to know the full sequence, right? We get the full sequence in one continuous range, ok? This is what we call the assembly problem.

You have something else called the mapping problem, right? As this is associated with the mapping, read mapping. So what is the mapping problem? So we got this bunch of reads, and you also have the reference genome sequence, ok? For example, let us say we sequence a human genome now. The reference genome sequence is available, so what we will do is take these reads and start to find out, ok, which part of the reference genome this read comes from. We are essentially mapping it back to the reference genome to know the location of the reads, and this will help us identify the changes and genetic mutations that are happening at those specific positions. So this is the mapping process; this cartoon shows you this, right?

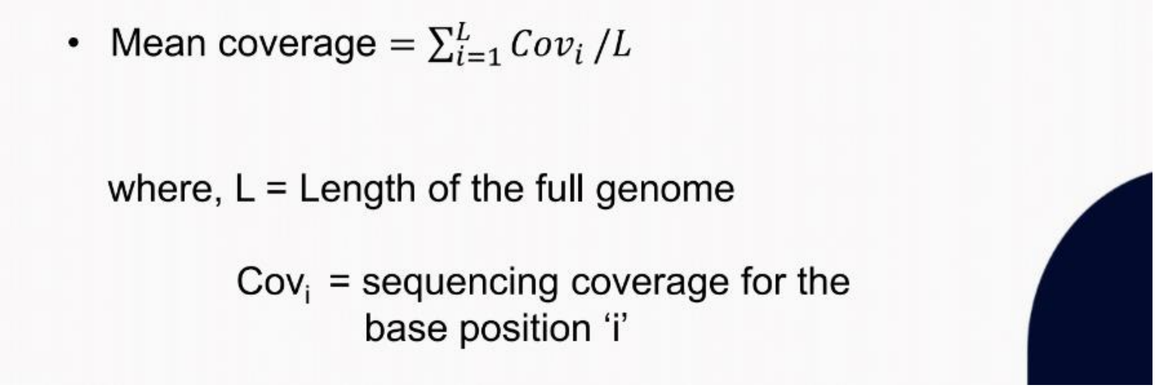
So for example, in the first read, we see, ok, this kind of matches with this part of the genome, so we put it there, right? We kind of keep track, ok, this reads maps here in this genome, and we then go on to the next read, right, and we see, ok, this one matches with this part of the genome, right, we put it there, ok. And we continue this process until we have mapped all that, ok, against the reference genome, ok. So this is kind of the mapping problem, in brief. So the term sequencing coverage that we use actually comes after we have done the mapping, ok?

So this is the sequencing coverage, or sometimes we also refer to this as sequencing depth, ok? So for each base in the reference sequence, the number of reads mapped to that position refers to the sequencing coverage, right? So this determines the sequencing coverage, ok? So I will just illustrate again with the cartoon, right? So after this mapping, if we take these four positions, right, position A, B, C, and D, we just check, right, how many reads are there that have been mapped to that position, ok?



Sequencing coverage or sequencing depth

So, for example, for position A, you have just one read mapping to that position, so you have coverage of one. For position B, you have four read mappings: coverage four, position C coverage five, and position D coverage two. So this is how we calculate coverage for each base in the reference genome sequence, ok? Now there is another term, which is called the mean coverage, ok? So this is the mean coverage for all positions across the whole genome, ok?



- Mean coverage = $\sum_{i=1}^{L} Cov_i\ /L$

where, L = Length of the full genome

Cov_i = sequencing coverage for the base position ‘i’

So if you take the coverage value for each position of the genome and the average, that will give you the mean coverage. So here is the mean coverage formula, right? So this is the coverage I summed over, which equals L, and then, of course, you have to divide by the length of the genome. So that is something that you have to do, right? So this is the length of the genome, ok?

So L is the length of the full genome, right, and coverage is the sequencing coverage for the base position, ok? Now, how do you represent this mean coverage? So when you look at genome sequencing projects, sometimes you will see this: 10x coverage. What does it mean? So the mean coverage is 10, ok? So on average, there are 10 reads mapping to each position of the genome. This term, on average, is very important, right?

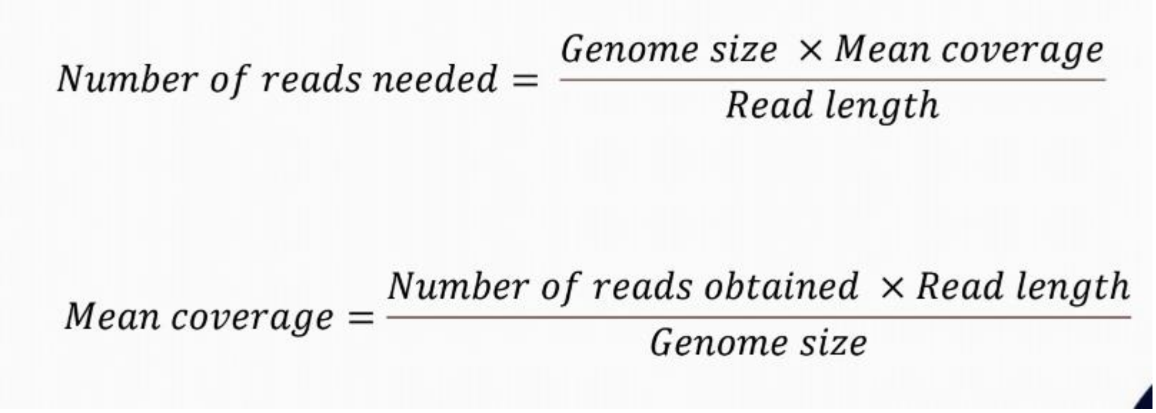
So it does not mean that all positions will have 10 reads mapped, ok? So this is something you need to be aware of, right? Even if you say, ok, mean coverage is 10x, right, it does not mean every position will have 10 reads, ok. So this is something that we observe across this data variation in coverage, ok? When you do a genome sequencing project, we do the mapping, and we see these fluctuations in coverage across the genome.

And this is observed because you have something called sampling bias and amplification bias, ok? So it can happen, right? In some positions, you have seen these examples: 1, 4, 7, and in some cases, it might be 0, right? In some positions, there might not be any read-mapped at all, ok? And then again, the coverage could fluctuate between 1, 5, 7, etc., ok? Again, this is something you need to be aware of, ok? Now, what is sampling bias? I mentioned sampling bias.

So this is actually because of the random selection of DNA fragments from the pool, right? So when you are preparing the library, you are fragmenting the genomic DNA, and you are generating these fragments, and only a part of this fragment is taken for amplification or direct sequencing. So there is a bias, right? Sometimes there could be a bias in which region, which part, or which fragments you are choosing for the next steps. So that will be the sampling bias. You can also have something called amplification bias, right?

So many of these next-generation methods, as we have discussed so far, require this amplification step. Okay, except nanopores, all of them require this amplification step, and this amplification can select certain DNA molecules. So it can preferentially amplify specific DNA molecules, right, and this happens because there are, I mean, a lot of factors, but one of them is the GC content, right? So there is an optimal GC content at which there is very good amplification; there could be a mismatch, so a mismatch will reduce amplification, etc. So there is a lot of amplification bias that will happen during this library preparation step, ok, and this is something that will again generate this variation in the coverage, ok. Now the question is: Why do we care about this mean coverage?

So this is very important. So one of the first points is that this has very important implications for designing an NGS experiment, and it helps in deciding which platform you want to use and the kind of throughput that you require. So this is the formula that we will use, right? When we are designing an experiment, the number of reads that are needed for our experiment is actually given by this. So if you have a, if you know the genome size and you know the coverage that you want, for example, 10x, 20x, or 30x, the mean coverage will be 10, 20, or 30, respectively, and divide by the read time, ok? So if you know the read length because you are going with a specific platform, you know the median read length or mid-read length. You can calculate how many reads you would need if you know the coverage. So again, that will help you choose the right kind of platform, right? Whether you need 100 million reads or 10 million reads, depending on that, you will choose the platform.



$$Number\ of\ reads\ needed = \frac{Genome\ size\ \times Mean\ coverage}{Read\ length}$$

$$Mean\ coverage = \frac{Number\ of\ reads\ obtained\ \times Read\ length}{Genome\ size}$$

And again, you can turn it around and actually calculate the mean coverage that you will get. So, you know the number of reads that you will get from a platform, and you know the read length of that platform divided by the genome size, right? So you can play around with this, right, and you can calculate which parameter you need, right, what kind of mean coverage or what kind of read number you need for your sequencing project. Now, there is a term called deep sequencing related to this coverage term. So this actually refers to sequencing a genome region multiple times, where the mean coverage would be 100, between 100 and 1000, or even more.

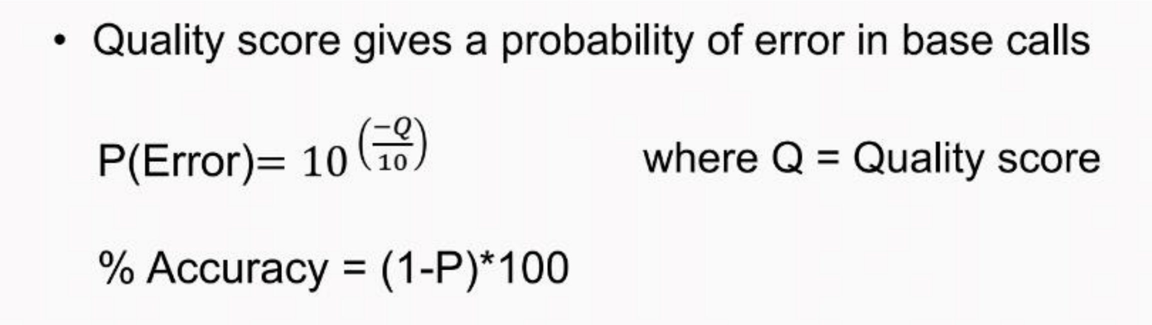
So it is really deep sequencing, right? So coverage is 100x, 1000x, or even more, right? So why do you do this deep sequencing? We want to identify rare clones or cell types within a cell population. For example, a cell type is present only in 1 percent of the cell population, right? So in that case, you probably want to go for, let's say, 1000 coverage or even more, right?

So this is kind of an application that we have, right, and the lower limit of detection frequency, right. So whether you will detect a cell type that is present at 1 percent frequency or 0.1 percent frequency is determined by the coverage of the deep sequencing experiments, right? Now coming to the second important term, right? This is a very important term, which is called the quality score. So this quality score actually is a measure of confidence in a base called by the detector from the signal, right?

So for each of these sequencing technologies, there is a detector; there is some sort of signal, right, by which the detector detects the bases or calls the bases. And this quality score defines the confidence in that base, right? And this is actually implemented in the instrument, and this is available for all bases that are sequenced in the instrument, ok? So this is something we may not be aware of, but we just see the end results, right?

So that is how the whole process happens. We may not know because it is already implemented in the instrument, but this is there, and the data is actually reported, as we will see when we talk about data formats. Again, this is something like measuring the signal-to-noise ratio. There is background noise, and how strong the signal is compared to the background noise kind of determines the quality score. And sometimes this quality score is also referred to as the PHRED score because of the first program that was used, called PHRED, for the Sanger sequencing-based quality termination. Now, how is this quality score important? So this quality score gives a probability of error in the base calls, ok?

So this is something; looking at the quality score, we can tell, ok, what is the chance that this base is wrongly called? This is not the right base, ok? So you see this formula here. So the probability of error is given by 10 to the power minus Q by 10, ok?



- Quality score gives a probability of error in base calls

P(Error)= $10^{\left(\frac{-Q}{10}\right)}$ where Q = Quality score

% Accuracy = (1-P)*100

So where Q is the quality score, ok? And then we can calculate this percentage accuracy, which is given by 1 minus P, the probability of error, into 100. So if you are calculating this accuracy in percentage, like we have seen for Illumina, right, 99.9 percent accuracy, or for other platforms, we have seen 90 percent, 95 percent, whatever type this is, we can calculate using this formula, right? So one of the things you will see again is something called Q10, Q20, Q30, Q40, ok? So what does this mean, right? So once you see this, you will see these terms for each platform, right?

So for example, in Illumina, you say, ok, we have like 99 percent of the base calls above Q30, which means they have 99 percent of the base calls above 30 quality score. So the error rate and probability of error are very low, and then you can calculate the accuracy, right? So you get this accuracy. For example, in Nanopore, you say, "Okay, we have achieved Q20."So most of the bases now are above this 20-quality score, ok?

So this is something that is very important. You see this term, ok, because this is used in determining the accuracy. So just to illustrate, if Q equals 10, then P equals the probability of error, which is 0.1, which means 1 into N errors. This is a high error rate.

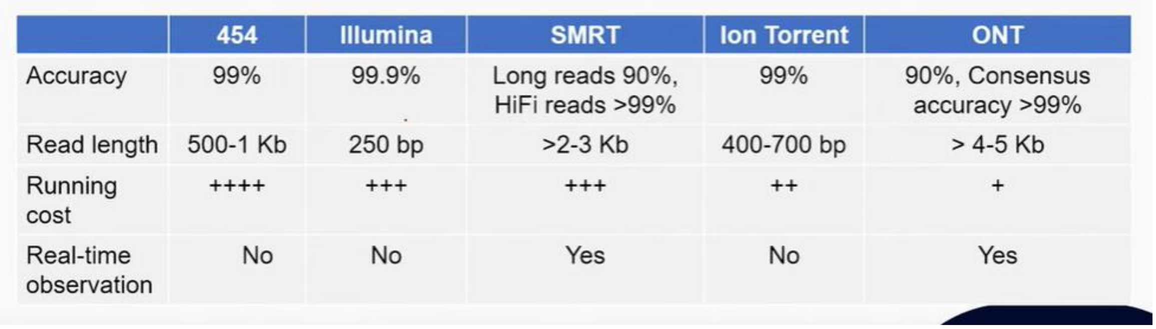
So you have 90 percent accuracy. If Q equals 40, then the probability of error is 10 to the power of minus 4, right? You just use this formula here, and this means 1 in 10,000 error, right? So the chance of a base being an erroneous call is actually very, very low, right? So a higher Q value means a lower chance of error, ok? Now one of the questions that you might have in your mind is: does this quality score apply to homopolymeric stretches? Would it accurately reflect the error that happens in homopolymeric stretches, for example, in 454 or in an ion torrent? So in a 454 or ion torrent, this length of homopolymeric stretch is determined from one single peak, right?

It is not like separate single peaks, separate signal peaks so you have these signal peaks coming for each base, right? So if you look at the quality score of these bases in the homopolymeric stretch, they are actually very close to each other, right? Then they do not differ. We learn so much from each other. So what researchers have seen is that the PHRED score is actually not informative enough, ok? So you cannot just rely on the PHRED score if you are looking at a homopolymeric stretch in 454 data or ion torrent data, right?

So, researchers have developed more complex base-calling models, right? I will not go into all those complex models, but what I am saying is that you need to actually be careful when just relying on the PHRED score in the case of homopolymeric stretches, because this detection happens from one signal peak, right? And if that peak is well resolved compared to noise, you might get a good quality score, ok? So, coming to the last part, right, so we have looked into coverage, we have looked into quality score, and the last part is about experiment design, right? Now, when we are starting a project or doing some experiments, one of the questions that we might have is, which NGS technology should I use for my experiment?

And this is a question that we always face, right, when we are doing our research. We also face this question, right, which NGS technology is suitable for me. Again, there is no single answer, right? It depends on the requirement, depends on your project, and depends on many factors. So, before we go there, I just wanted to give you an overview of this comparison of different NGS technologies that we have discussed so far. We have 454, Illumina, SMRT, Ion Torrent, and ONT, ok? And there are a few important points that are really relevant for us.

One is accuracy, right? This is something that we care about. So 454 gives you 99 percent accuracy, right? Illumina is about 99.9 percent, SMRT Long Beach is about 90 percent, let us say, and HiFi can get above 99 percent. Ion Torrent has about 99 percent accuracy, and Oxford Nanopore has about 90 percent accuracy, but if you are using consensus calls, you will get above 99 percent. For read length, right, so 454 we have talked about 500 to 1 kB, Illumina is about 250 base pairs, if you are doing paired end you get 2 times 250, SMRT you will get above 2 3 kB, right, Ion Torrent is about 400 to 700 base pairs and ONT is more than 4 to 5 kB.



	454	Illumina	SMRT	Ion Torrent	ONT
Accuracy	99%	99.9%	Long reads 90%, HiFi reads >99%	99%	90%, Consensus accuracy >99%
Read length	500-1 Kb	250 bp	>2-3 Kb	400-700 bp	> 4-5 Kb
Running cost	++++	+++	+++	++	+
Real-time observation	No	No	Yes	No	Yes

Running cost is a very important point, right? With the funding that you have, So the running cost will be $454, which was very expensive, so it is not available anymore. Illumina SMRT's running cost is on the higher side because of the polymerase and the modified nucleotides that you use. Ion Torrent is slightly more economical compared to Illumina SMRT because you do not use any modified nucleotides. And of course, since we have discussed all the methods, ONT is the most cost-effective because it is doing direct sequencing with no polymerase or dNTP involved.

Coming to real-time observation, so 4 5 4 Illumina Ion Torrent are not real-time observations, whereas SMRT and ONT are real-time observations, right? So again, if real-time observation is important for your experiment, you want to consider SMRT and ONT. Throughput again coming to all these platforms, now we can adjust this throughput because we have these different chips or different kinds of sequence sequences, right, to cater to different throughputs. Now there are a few things that we need to first consider before we can answer that question, right? Which NGS platform is most suitable for my experiment? So the first two points are, of course, the type of experiment that you are doing and the end goal of that experiment, right?

Then, of course, what is the read length that you need, the accuracy that you need, and the coverage that you require, and of course, the funds available, right, how much money do you have? So all these things you need to carefully think about before you can decide on the sequencing platform, ok? Types of experiments: these are the types of experiments that can be run that you can run, right? So you can be doing genome sequencing or transcriptome analysis; you can look at epigenetic modifications or metagenomic studies; or you can do something called amplicon sequencing which is sequencing PCR products. You have only a small region of the genome that you are interested in; you are just amplifying that out by PCR and sequencing that region.

So this is called amplicon sequencing, ok? Now again, for each of these types of experiments, you can have different end goals, ok? So for example, for genome sequencing, you have de novo assembly, and you might be interested in sequencing a new genome, right? You can have identification of SNPs and indels, right? You are studying a population and looking at genetic variation, for example, or you might be interested in structural variations, right? recombination, rearrangements, infusions, etc. Again, for transcriptome sequencing, the most common goal is to measure the expression levels of genes, right? So, this is the one common goal, but then you can have other goals also, right?

For example, you might want to study alternative splicing events, or you might be interested in looking at the differences in expression of different alleles of the same gene. So, in diploid organisms, we have these two different alleles, let us say of the same gene, and you want to see how differently they are expressed, whether they are expressed at the same level or there is some difference in expression of those alleles, right? So these are the kinds of studies you might be doing as well. Again, in epigenomics, you have methylation patterns; you can look at histone modifications or transcription factor binding; these are different aspects. And one of the key questions for all of these projects, right, is whether your experiment is genome-wide or targeted to specific regions of the genome, because this will again determine which platform you use, whether you use Illumina, whether you use wrong-link sequencing, etc.

, right. So, these are the questions; these are the aspects that you need to think about carefully, right? And I will give you some examples, right, and that will help you out with this process, right? So, this is the first example of how you choose the best experimental design for your project, right? So let us take this scenario where what is required is that we want to do a de novo assembly of a genome from a new organism that has not been sequenced before. And we want all the information on single nuclear polymorphisms, insertions, deletions, and structural variants for comparison with a related species.

So we want to do the genome sequencing, and we also accurately want to determine these differences compared to a related species. So here the assembly problem is there, but we also want to add this mutation detection problem, right? So you need long-read sequencing. So what would be the solution here? So the solution is that you can probably use long-read sequencing plus short-read sequencing. You can use this combination of both because long-read, which may not be like traditional long-read methods, which may not be very accurate, but they can give you structural variant information, and they can help in the genome assembly process. Whereas short-reads are much more accurate, as we have seen, they will be more useful for this single molecular polymorphism study or indel study, right?

So this is something; this is one solution. The other solution, if you can, if you have access to high-accuracy long-reads sequencing, is, as we have talked about, for example, Hi-Fi in SMRT or in consensus sequencing in nanopore technology. So, this is the solution. Let us take another example, right? So, choosing the most appropriate experiment design is right. So here in this thought experiment, right, we want to do a deep sequencing of a 1 kB region, ok, and why do you want to do deep sequencing? You know what deep sequencing is: coverage of 1000 or maybe more of this region for the identification of rare variants.

For example, we are looking at a very important gene, right, in humans, and these variations in this gene can lead to certain diseases or affect certain metabolic metabolites, right, etcetera that can lead to disease. So, this is something we want to study. Now the answer here is not really straightforward, ok? So, one of the questions that you might have that you want to ask still is: Where do you expect these variations to occur?

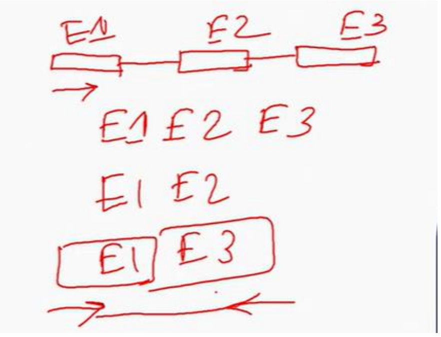
These are variations, so I will draw them here. So, are these variations near the end of this 1 kB region, ok? Do we expect these rare variants to occur only between the last 100 base pairs here and the first 100 base pairs here, right, of this region, or do you expect these variants to occur throughout the genome, throughout this 1 kB region? So not just at the end, but throughout this 1 kB fragment, ok. So, these two cases will have two different solutions, ok?

It is not just one solution. So, in the first case here, this one, right, since they are at the end, right, end of this region, right, and the start of this region, right. So you can actually reach these regions with a short read sequence, ok? And perhaps then you can do something called paired-end, right? We have discussed this paired-end sequencing. You can do paired-end sequencing perhaps on the Illumina platform; maybe this is a solution, ok?

So this might be one solution, but then again, you might have to be careful about this fragment size. So Illumina might be able to take only certain fragment sizes for this cluster generation or bridge amplification, ok? So this is something we need to inquire about or consult with the literature and the Illumina sequencing platform. In the second case, you cannot go ahead with this paired-end sequencing because these mutations are scattered throughout this 1 kB region, ok?

So what you have to do is either go through this long read sequencing, right? So either use the nanopore platform or SMRT sequencing and use the consensus call, right? CCS reads in SMRT or the consensus call in nanopore sequencing, right? So, as you can see, this requires a lot of thought, right? First, you need to know your requirements, and then you can think about what kind of solution would be most appropriate for you and what these technologies are capable of. So the third and final example, right? So here we want to do genome-wide quantification of gene expression, and we want to discover splice variants, ok?

This is a very interesting example, right? So let us say we are working with a species that has around 5000 genes, ok, and we want to measure the expression level of all these genes and also look at splice variants, right? So you understand what splice variants are, right? So we have this exon-intron structure, right, and they can combine in any combination, right? So the splice variants are E1, E2, E3, right, this is combining all exons you can have E1, E2 only or E1, E3, right, so whether you get E1, E3 or not.



So how do you actually find out, right, whether this is? So here just maybe doing the single end sequencing may not be good enough and probably you want to do paired end sequencing, right, because in paired end if you see, for example, in this scenario if you have an E1, E3 combination and you get these reads, one of which falls on E1, the other one falls on E3, and they are from the same DNA fragment, right, then it will tell you, ok, this is probably a splice variant, right, this is an E1, E3 splice variant, this is a different combination than having and having all these exons in there, ok. So that concludes our discussion. These are the references that we have used, and just to conclude, right, so we said we showed, so we discussed that right sequencing coverage is an important metric for choosing the right NGS platform. Right, you can play around with read length, coverage, etcetera, and this will help you decide on the right platform. The quality score we have discussed gives us an expected error rate, which could be used for downstream analysis. Right now, we will incorporate these quality scores later on in subsequent classes and see how that is important. And finally, we talked about NGS experiment design. So designing an NGS experiment requires careful consideration of many features, and depending on your needs and availability, you will find the optimal solution. Thank you.